GENERALIZED LINEAR MODELS
Lecture Notes for BST/STA 223

UC Davis

Winter 2021

1 Preliminaries

1.1 Regression Models and Methods

Predictors (independent variables, covariates, features) $X \mapsto$ Response (dependent variable) Y

Goals:

- *Establish* whether a relationship exists (testing for regression, *Inference*)
- Develop appropriate model to *describe* relationship between predictors and responses (*Description, Predictor Selection, Modeling Interactions, Interpretation, Findings*)
- *Predict* responses for new cases (*Prediction, Classification*)

Types of variables

	Dependent			
Independent	binary	count	multinomial	continuous
binary(0/1)	2×2 table or binary reg.	Poisson regr.	Multinomial regr.	t-test
nominal	log-linear or binary regr.	Poisson regr.	Multinomial regr.	ANOVA
continuous	binary regr.	Poisson regr.	Multinomial regr.	Multiple regr.

Note: Multinomial variables with K outcomes or levels are represented by $K-1$ "dummy" or binary variables, usually referred to as *indicators*.
Important: Define a good baseline (where all indicators=0)

We aim at general unifying theory and methodology to cover all these various regression relations under one model → the generalized linear model (GLM).
Important cases: Dependent variable is a binary outcome or a count (in these cases the variance is expected to be a function of the mean), or is a Gaussian response.

General quantification of regression: Starting point is joint distribution $\mathcal{L}(X, Y)$
Regression is $E(Y|X = x) = m(x)$

More general approach: Conditional distribution $\mathcal{L}(Y|X=x) \equiv P(Y \le y|X=x)$ for all x and y.

1.2 Notation

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \in \mathcal{R}^n, \;\; \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathcal{R}^n$$

will denote n-vectors, where $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$ is a p-vector of covariates.

For $f : \mathcal{R}^p \to \mathcal{R}$, let $\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x_0}) = \begin{pmatrix} \partial f/\partial x_1 \\ \vdots \\ \partial f/\partial x_p \end{pmatrix}_{\mathbf{x}=\mathbf{x_0}}$, the *score vector* or *gradient* at $\mathbf{x_0}$

$$\frac{\partial^2 f}{\partial \mathbf{x}^2}(\mathbf{x_0}) = (\frac{\partial^2 f}{\partial x_i \partial x_j})_{\mathbf{x}=\mathbf{x_0}}, \;\; 1 \le i,j \le p, \quad \text{the } \textit{Hessian matrix} \text{ at } \mathbf{x_0}$$

$$\det(\frac{\partial^2 f}{\partial \mathbf{x}^2}) = \textit{Jacobian}\ \ (\textit{determinant})$$

For $f : \mathcal{R} \to \mathcal{R}$, $\mathbf{x}, \mathbf{y} \in \mathcal{R}^n$, we write short-hand

$$f(\mathbf{x}) = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \mathbf{y} \in \mathcal{R}^n$$

Example: $\mathbf{x}^\lambda$, where $\mathbf{x} \in \mathcal{R}^n$, $\lambda \in \mathcal{R}$, stands for $\begin{bmatrix} x_1^\lambda \\ \vdots \\ x_n^\lambda \end{bmatrix} \in \mathcal{R}^n$

Let $\mathbf{X} \in \mathcal{R}^p$ be a r.v. with pdf $f(\mathbf{x}, \theta)$, where θ is a parameter. Given a function $h : \mathcal{R}^p \mapsto \mathcal{R}$,

$$E_\theta \left[h(\mathbf{X}) \right] = \int h(\mathbf{x}) f(\mathbf{x}, \theta) \, d\mathbf{x}$$

1.3 Modelling

GLM is about modelling complex as well as simple regression relations. We are given observations (x_i, y_i), $i = 1, \ldots, n$, with univariate or multivariate predictors (covariates) $x_i \in \mathcal{R}^p$, $p \geq 1$, and responses $y_i \in \mathcal{R}$ (usually). There are two main purposes of regression and also of GLM:

- Significance of regression and describing the nature of the regression relation (linear, quadratic, saturation etc.)
- Prediction and classification of responses given new predictor levels

The data obtained for n different subjects or experimental units (index $i = 1, \ldots, n$) are assumed to be independent. We usually consider the predictor vectors to be fixed, i.e., non-random.

1.4 Preliminaries on likelihood

Assume $y_i \sim_{indep.} f(y, \theta_i)$, $\theta_i = \theta(x_i)$, $i = 1, \ldots, n$.

Let $\mu = Ey$ (i.e., $\mu_i = Ey_i$). Then the log-likelihood of y expressed as a function of $\mu = \mu(\theta)$ is

$$l(\mu, y) = \sum_{i=1}^{n} \log f(y_i, \theta_i),$$

where $\mu_i = \int y f(y, \theta_i) dy$, $\mu_i = h^{-1}(\theta_i)$, $\theta_i = h(\mu_i)$ for some invertible function h (that is assumed to exist). Here $f(\cdot, \theta_i)$ is the pdf or pmf of y_i, given parameters $\theta_i = \theta(x_i)$.

Example: y_i normal with known variance σ^2,

$$\mu_i = \beta_0 + \beta_1 x_i, \;\; \theta_i = \mu_i, \;\; h \equiv \text{id}$$

$$f(y, \theta_i) = \frac{1}{\sqrt{2\pi}\sigma} \;\; \exp(-\frac{(y-\theta_i)^2}{2\sigma^2})$$

1.5 Deviance

A quantification of the deviation of the model log likelihood from a null likelihood. In the null likelihood, the parameters μ correspond to actual observations y:

$$D^*(y, \mu) = 2 \;\; \{l(y, y) - l(\mu, y)\}$$

Example: Gaussian model with known variance σ^2, $y_i \sim N(\mu_i, \sigma^2)$,

$$f(y_i, \mu_i) = (2\pi\sigma^2)^{-1/2} \exp\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\}$$

$$l(\mu, y) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2\sigma^2}$$

$$l(y, y) = -\frac{n}{2}\log(2\pi\sigma^2), \qquad \text{therefore}$$

$$D^*(y, \mu) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma^2}$$

In this case the deviance corresponds to the usual residual sum of squares, scaled by σ^2.

2 Likelihood

2.1 Estimating Equations and Information

Certain relations hold for the derivatives of the log-likelihood function l; these are basic for the properties of the likelihood estimates (MLEs).

Assume $y \sim f(\cdot, \theta)$, a pdf with parameter $\theta \in \mathcal{R}$. Then

$$\int f(y,\theta)dy = 1, \quad \text{for all } \theta \in \Theta.$$

Here Θ is a compact parameter space (e.g., an interval). If we can exchange integration and differentiation,

$$\begin{aligned} 0 &= \frac{d}{d\theta}\int f(y,\theta)dy = \int \frac{d}{d\theta} f(y,\theta)dy \\ &= \int \frac{d}{d\theta}\log f(y,\theta) f(y,\theta)dy = E_\theta(\frac{\partial l}{\partial \theta}) \end{aligned} \tag{2.1}$$

When can we make the exchange? It requires existence of a dominating function $\psi(\cdot)$: If $\sup_\theta |\frac{d}{d\theta} f(y,\theta)| \leq \psi(y)$, where $\int \psi(y)dy < \infty$, then the exchange is possible. It is enough that $f(y,\theta)$ is locally dominated integrable: For each $\theta \in \Theta$, there exists a neighborhood $U(\theta)$ and a local dominating function $\psi_\theta(\cdot)$, s.t. for all $\zeta \in U(\theta)$,

$$\sup_{\zeta \in U(\theta)} |\frac{d}{d\tilde\theta} f(y,\tilde\theta)|_{\tilde\theta=\zeta}| \leq \psi_\theta(y), \quad \int \psi_\theta(y)dy < \infty.$$

Assuming there is a dominating function, and extending (2.1) to the second derivative,

$$\begin{aligned} 0 &= \frac{d^2}{d\theta^2}\int f(y,\theta)dy = (\text{exchange}) \int \frac{d^2}{d\theta^2} f(y,\theta)dy \\ &= (\text{expand}) \int \frac{(\frac{d^2}{d\theta^2} f(y,\theta))f(y,\theta) - (\frac{d}{d\theta} f(y,\theta))^2}{f^2(y,\theta)} f(y,\theta)dy + \int \frac{(\frac{d}{d\theta} f(y,\theta))^2}{f^2(y,\theta)} f(y,\theta)dy \\ &= \int \frac{d^2}{d\theta^2} l(y,\theta) f(y,\theta)dy + E_\theta(\frac{d}{d\theta} l(Y,\theta))^2 \\ &= E_\theta(\frac{\partial^2 l}{\partial \theta^2}) + \text{var}_\theta(\frac{\partial l}{\partial \theta}), \\ & \quad \text{since} \quad E_\theta(\frac{\partial l}{\partial \theta}) = 0 \quad \text{by (2.1)}. \end{aligned}$$

We find

$$E_\theta(\frac{\partial l}{\partial\theta}) = 0 \quad \text{(score equation)} \tag{2.2}$$

$$\text{var}_\theta(\frac{\partial l}{\partial\theta}) = -E_\theta(\frac{\partial^2 l}{\partial\theta^2}) \quad \text{(information equation)} \tag{2.3}$$

and for the third derivative:

$$E_\theta(\frac{\partial^3 l}{\partial\theta^3}) + 3\text{cov}_\theta(\frac{\partial^2 l}{\partial\theta^2}, \frac{\partial l}{\partial\theta}) + E_\theta(\frac{\partial l}{\partial\theta})^3 = 0. \tag{2.4}$$

(2.2)-(2.4) are known as "Bartlett's identities".

For the *score vector* or *score statistic* $U = \frac{\partial l}{\partial\theta}$, (2.1) implies $E_\theta U = 0$, motivating the *Estimating Equation*

$$U(\theta, y) = 0 \tag{2.5}$$

with solution (if it exists and is unique) $\hat{\theta}$.

The *Fisher Information* is $\quad I(\theta) = -E_\theta(\frac{\partial^2 l}{\partial\theta^2})$,
so that by (2.3), $I(\theta) = \text{var}_\theta(\frac{\partial l}{\partial\theta}) = \text{var}_\theta(U(\theta, Y))$.
Note that the p-dimensional case where $\theta \in \mathcal{R}^p$, $p \geq 1$, is always included.

For $p > 1$, score and Fisher information become

$$U(\theta_0) = U(\theta_0, y) = \begin{pmatrix} \partial \log f(y,\theta)/\partial\theta_1 \\ \vdots \\ \partial \log f(y,\theta)/\partial\theta_p \end{pmatrix}_{|\theta=\theta_0} \qquad p \times 1 - \text{score vector}$$

$$I(\theta_0) = \left(-E_{\theta_0} \frac{\partial^2}{\partial\theta_k \partial\theta_l} \log f(y,\theta)|_{\theta=\theta_0}\right)_{1 \leq k,l \leq p} \qquad p \times p - \text{information matrix}$$

2.2 Review of asymptotics

1. *Asymptotic normality of score*

For a sample of independent observations, we have

$$f(y,\theta)=\prod_{i=1}^{n} f(y_i,\theta_i), \quad y=\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

so that $\log f(y,\theta)=\sum_{i=1}^{n}\log f(y_i,\theta_i) \sim$ sum of n independent random variables, as $n\to\infty$. Standardizing by $I(\theta_0)^{-1/2}$ allows to apply the CLT. More specifically, as

$$\begin{aligned} U(\theta_0,y) &= \frac{\partial}{\partial\theta}[\log f(\theta,y)]_{\theta=\theta_0}=\sum_{i=1}^{n}\left[\frac{\frac{\partial}{\partial\theta}f(y_i,\theta)|_{\theta=\theta_0}}{f(y_i,\theta)|_{\theta=\theta_0}}\right], \\ I(\theta_0) &= \operatorname{cov}(U(\theta_0,Y))=\left[\operatorname{cov}\left(\sum_{i=1}^{n}\frac{\partial}{\partial\theta_j}\log f(y_i;\theta)|_{\theta=\theta_0}\,,\,\sum_{i=1}^{n}\frac{\partial}{\partial\theta_k}\log f(y_i;\theta)|_{\theta=\theta_0}\right)\right]_{1\le j,k\le p} \end{aligned}$$

one finds

$$[I(\theta_0)]^{-1/2}U(\theta_0,Y)\to_D N(0,I_p) \tag{2.6}$$

where $I_p=\begin{bmatrix} 1 & \cdots & 0 \\ 0 & \cdots & 1 \end{bmatrix}_{p\times p}$ is the $p\times p$ identity matrix, and $\to_D$ denotes convergence in distribution. Here $\theta_0\in\mathcal{R}^p$ is the true underlying parameter value. This yields the *Score Statistics* and *Score Test* for the null hypothesis $H_0:\theta=\theta_0$.

2. *Asymptotic normality for the MLEs*

Estimating Equation leads to estimates $\hat\theta: \quad U(\hat\theta,y)=0.$ Under suitable assumptions,

$$[I(\theta_0)]^{1/2}(\hat\theta-\theta_0)\to_D N(0,I_p) \tag{2.7}$$

This yields the *Wald Statistic* and *Wald Test*. When applying (2.7) (and also (2.6)), $I(\theta_0)$ is replaced by $I(\hat{\theta})$ which is justified by *Slutsky's theorem*.

3. *Asymptotic distribution for the log-likelihood*

   By Taylor expansion, from (2.7),

$$2\{l(\hat{\theta}, y) - l(\theta_0, y)\} \to_D \chi^2_p \tag{2.8}$$

   where $l(\hat{\theta}, y)$ stands for the log-likelihood at the observed MLE. This yields the *Likelihood Ratio Statistic* and *Likelihood Ratio Test*.

Asymptotic inference can be based on either of these approximations; depending on which one is used, the results may differ.

Examples

1. A simultaneous confidence region for θ_0:

$$\{\theta : \quad 2\{l(\hat{\theta}, y) - l(\theta, y)\} \leq \chi^2_p(1-\alpha)\}$$

   level is $(1-\alpha)$, based on (2.8).

2. A confidence interval for θ_{0k} (the k-th component of the vector θ_0):

$$\hat{\theta}_k \pm \Phi^{-1}(1 - \frac{\alpha}{2})[I(\hat{\theta})^{kk}]^{1/2},$$

   at level $(1-\alpha)$, $I(\hat{\theta})^{kk}$ denoting the k-th diagonal element of $[I(\hat{\theta})]^{-1}$ (a $p \times p$-matrix, $1 \leq k \leq p$), obtained from (2.7) (Wald statistic), by projecting on the k-th component.

   Note: This device can be used for one- or two-sided confidence regions and tests. Multiple testing can be a problem (Bonferroni or False Discovery Rate adjustments).

3. Testing H_0: $A\theta = \zeta$ (general linear hypothesis),
   for a $q \times p$ matrix A of rank q, $q \leq p$, $\zeta \in \mathcal{R}^q$. Consider the MLE $\hat{\theta}$, $\theta \in \mathcal{R}^p$.

Setting $B = (AI^{-1}A^{\top})^{-1/2}AI^{-1/2}$, a $q \times p$ matrix, one finds $BB^{\top} = I_q$.

Then, from the asymptotic normality of MLEs (2.7),

$$BI^{1/2}(\hat{\theta} - \theta_0) \rightarrow_D N(0, I_q) \tag{2.9}$$

leading to the Wald-type test statistic

$$[(AI^{-1}(\theta)A^{\top})^{-1/2}A(\hat{\theta} - \theta)]^{\top}[(AI^{-1}(\theta)A^{\top})^{-1/2}A(\hat{\theta} - \theta)] \rightarrow_D \chi^2_q \tag{2.10}$$

so that under H_0, using $A\theta = \zeta$,

$$[(AI^{-1}(\hat{\theta})A^{\top})^{-1/2}(A\hat{\theta} - \zeta)]^{\top}[(AI^{-1}(\hat{\theta})A^{\top})^{-1/2}(A\hat{\theta} - \zeta)] \rightarrow_D \chi^2_q \tag{2.11}$$

Here Slutsky's theorem justifies to replace $I(\theta)$ by $I(\hat{\theta})$. This is known as the *Sandwich Formula.*

2.3 Simultaneous inference

To obtain simultaneous (also referred to as joint) inference for a parameter vector $\beta \in \Re^p$, a common starting point is asymptotic normality of corresponding estimates $\hat{\beta}$:

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_D N(0, \tilde{\Sigma})$$

where $\tilde{\Sigma}$ is the asymptotic covariance matrix. The following is general and includes MLEs as a special case. For the special case of MLEs, $I_\infty = \lim_{n\to\infty} I(\beta)/n$ and $\tilde{\Sigma} = I_\infty^{-1}$, where I_∞ is the asymptotic limiting information matrix (which is assumed to exist), a $p \times p$-matrix. The asymptotic distribution result motivates

$$\hat{\beta} \sim N(\beta, \Sigma), \quad \text{approximately, for some matrix} \quad \Sigma.$$

For the case of MLEs, $\Sigma = I^{-1}(\hat{\beta})$ where $I(\hat{\beta})$ is the finitely estimated information matrix. Then

$$(\hat{\beta} - \beta)^\top \Sigma^{-1} (\hat{\beta} - \beta) \sim \chi^2_p, \quad \text{approximately.} \tag{2.12}$$

From this obtain the simultaneous $(1 - \alpha)$ confidence region:

$$C_{1-\alpha} = \{\beta : \quad (\hat{\beta} - \beta)^\top \Sigma^{-1} (\hat{\beta} - \beta) \leq \chi^2_p(1 - \alpha)\}.$$

Since Σ is nonnegative definite symmetric, there exists an orthonormal matrix Q s.t.

$$Q^\top Q = I_p, \quad Q\Sigma^{-1}Q^\top = \Lambda, \quad \Sigma^{-1} = Q^\top \Lambda Q,$$

I_p being the p-dimensional identity matrix and Λ a diagonal matrix.
Defining vectors $\gamma = Q(\hat{\beta} - \beta)$, obtain

$$(\hat{\beta} - \beta)^\top \Sigma^{-1} (\hat{\beta} - \beta) = \gamma^\top \Lambda \gamma = \sum_{j=1}^{p} \lambda_j \gamma_j^2, \tag{2.13}$$

where λ_j are the diagonal elements of Λ and also the eigenvalues of Σ^{-1}. The j-th row of Q is the eigenvector e_j of Σ^{-1}.

Note that Q can be interpreted as a rotation that maps the unit vectors $\tilde{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ with the 1 in j-th position to e_j, and $Q^T = Q^{-1}$ is the inverse rotation. Therefore, the geometrical object for which

$$\sum_{j=1}^{p} \lambda_j \gamma_j^2 \leq \chi^2_p(1 - \alpha)$$

is an ellipsoid.
The directions of the main axes of this $(1 - \alpha)$ ellipsoid are given by the e_j, and their half lengths by $\sqrt{\chi^2_p(1 - \alpha)/\lambda_j} = \sqrt{\chi^2_p(1 - \alpha)\tilde{\lambda}_j}$, where $\tilde{\lambda}_j, e_j$ are eigenvalues and eigenvectors of the matrix Σ, which thus determines the $(1-\alpha)$ confidence region (c.r.) $C_{1-\alpha}$. These regions can then be used to test complex hypotheses about $\beta = (\beta_1, \ldots, \beta_p)$, e.g., $H_0 : \ \zeta^T \beta = \rho$

(linear hypothesis) or $H_0: \ \beta_1\beta_2 > 0$ for $p = 2$, i.e., β_1, β_2 have the same sign.

Always keep in mind the distinction between a *statistical null hypothesis* and a *scientific hypothesis* to be established, were typically the latter will correspond to the alternative hypothesis.

3 Smoothing Methods

3.1 Basics

Given: Scatterplot data $(x_i, y_i), \quad i = 1, \ldots, n$

Goal is to estimate $g(x) = E(y|X = x)$, the regression function, without assuming a parametric form.

Important distinction: *Random Design* (observational studies, the x_i are considered random) versus *Fixed Design* or non-random design (experimental studies, the x_i are considered non-random). In the fixed design case, write alternatively

$$y_i = g(x_i) + \varepsilon_i, \quad \varepsilon_i \sim_{i.i.d.} (0, \sigma^2).$$

Assumption: g is "smooth", at least continuous. Common assumption is $g \in \mathcal{C}^2$ (twice continuously differentiable) on its domain.

There are various nonparametric smoothing methods available for estimating g.

Cubic smoothing splines

Find locally based cubic polynomials $g_\lambda(\cdot)$ which minimize

$$\sum (y_i - g(x_i))^2 + \lambda \int g^{(2)}(x)^2 dx \to \min \tag{3.1}$$

w.r. to $g \in \mathcal{C}^2$. The second term is a *smoothness penalty* and λ is a *smoothing* or *tuning* parameter. This is an example of a *penalized least squares problem*, which can be translated into a linear minimization problem. The solution and thus the smoothing spline fit depends on smoothing parameter λ, which determines the *trade-off* between smoothness and closeness to the data, formally: between *bias* and *variance* of the curve estimator.

$$\begin{cases} \lambda \uparrow \infty : & \hat{g}_\lambda \to \text{simple linear regression fit.} \\ \lambda \downarrow 0 : & \hat{g}_\lambda \to \text{interpolation of data.} \end{cases}$$

Kernel estimators

The smoothing parameter is the bandwidth h, with smoothing window $[x-h, x+h]$ if smoothing at x with a kernel of domain $[-1,1]$. Let $S_{(i)} = (X_{(i)} + X_{(i-1)})/2$, where $X_{(i)}$ is the i-th order statistic of the X_i, and let $y_{[i]}$ denote the concomitant of $X_{(i)}$.

(a) Convolution type

$$\hat{g}_C(x) = \sum_{i=1}^{n} y_{[i]} \int_{S_{(i)}}^{S_{(i+1)}} \frac{1}{h} K(\frac{x-u}{h}) du \tag{3.2}$$

(b) Quotient type (Nadaraya-Watson 1964)

$$\hat{g}_Q(x) = \sum_{i=1}^{n} \frac{1}{h} K(\frac{x-X_i}{h}) y_i / \sum_{i=1}^{n} \frac{1}{h} K(\frac{x-X_i}{h}) \tag{3.3}$$

(c) Local linear type (Gram 1879)

$$\hat{g}_{LL}(x) = \hat{\beta}_0(x) \quad \text{where} \;\; (\hat{\beta}_0(x), \hat{\beta}_1(x)) \;\; \text{are the minimizers of}$$

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1(X_i - x)))^2 K(\frac{x-X_i}{h}) \tag{3.4}$$

Kernel K is usually chosen as a symmetric, positive density function. Examples

$$K(x) = \frac{3}{4}(1-x^2)|_{[-1,1]} \quad \text{(Epanechnikov),}$$

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad \text{(Gaussian),}$$

$$K(x) = \frac{1}{2} 1_{[-1,1]} \qquad \text{(Rectangular)}$$

Rectangular and Epanechnikov kernels have compact domain $[-1, 1]$.

The local linear estimates can be analogously extended to the case of local polynomial fitting, and also to the estimation of derivatives. If estimating the $\nu-$th derivative, the order of the locally fitted polynomial should be $\nu + k$, where $k = 1, 3$ or sometimes a larger odd number. Example: Derivatives of growth curves are growth velocities, often much more useful than growth curves which are monotone increasing.

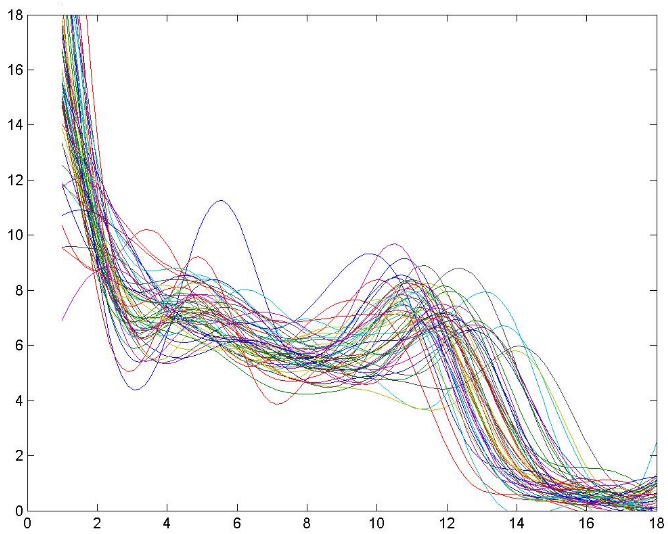


Figure 1: Estimated growth velocities from discrete (yearly) height measurements with some noise (corresponds to fixed design case) for $n = 54$ girls from the Berkeley growth study. The pubertal growth spurt and also an additional mid growth spurt are clearly visible.

Any practical implementation of a kernel or other smoother is based on input in the form of a scatterplot (x_i, y_i), $i = 1, \dots, n$, on a tuning parameter h and on a grid of output points x where one would like to see the estimate (usually chosen as a dense regular grid on the domain of interest, and unrelated to the values of the x_i.) Fixing just one output point x, a shortcut notation to denote the output of the smoother using the above ingredients is $S(x; (x_i, y_i)_{i=1,\dots,n}; h)$.

3.2 Additional Topics

Bandwidth choice methods/tuning parameter selection

Do not rely on an automatic method, always check visually whether the chosen bandwidth/tuning parameter makes sense.

Cross-validation (CV), based on minimizing the sum of one-leave-out squared prediction errors $CV(h) = \sum (y_i - \hat{g}^{(-i)}(x_i))^2$, where $\hat{g}^{(-i)}$ is the estimate of g constructed from a reduced sample where (x_i, y_i) is omitted; occasionally may lead to severe undersmoothing.

m-fold cross-validation, dividing the sample into m subsamples, leaving each out iteratively.

Generalized Cross-validation (GCV): a faster simplified version of CV.

Plug-in Methods: are based on estimating the Mean Squared Error or Integrated Mean Squared Error then minimizing w.r. to the bandwidth.

Subjective Choice: Start with a very small bandwidth that is undersmoothing (random wiggles are present). Then increase it successively by an increment on the log scale, i.e., $h_{k+1} = \rho h_k$, where for example $\rho = 1.2$. Look at the ensemble of fitted curves (overlay) and stop when oversmoothing occurs (features are missed, can use the runs test to determine when oversmoothing occurs).

Beware of *oversmoothing* (small variance, large bias) and of *undersmoothing* (large variance, small bias), where the latter is also known as overfitting. The bandwidth selection defines the necessary compromise between variance and bias. Runs tests can be used to diagnose oversmoothing.

B- and P-Splines

These are piecewise polynomials that are defined on subintervals and are smoothly connected at the endpoints of these intervals. The points separating subintervals are *knots*. The coefficients of these polynomials are usually fitted by least squares for the case of B-splines, whence these splines are also called regression splines.

Once the knots are selected (default is equidistant, as non-equidistant knot placement leads to difficult minimization problems), this determines a B-spline basis $B_1, B_2, \dots B_K$ of functions on a given support interval $\mathcal{T}$, where $K = m + 4$ if a cubic spline with m interior knots is used. These functions form a basis in a function space of smooth square integrable functions. An example for a cubic spline on domain $[0, 1]$ with interior knots $0 < \xi_1 < \xi_2 < \dots < \xi_m < 1$ is the basis

$$\{B_1, \dots, B_K\} \equiv \{1, t, t^2, t^3, (t-\xi_1)^3 1_{\{t\geq\xi_1\}}, (t-\xi_2)^3 1_{\{t\geq\xi_2\}}, \dots, (t-\xi_m)^3 1_{\{t\geq\xi_m\}}\}$$

Note: These basis functions can be linearly transformed to form an *orthonormal* set of basis functions $\tilde{B}_j$ by Gram-Schmidt orthonormalization, which means that

$$\int \tilde{B}_j(t)\tilde{B}_k(t)dt = \delta_{jk}, \quad \delta_{jk} = 0 \quad \text{if} \quad j \neq k, \quad \delta_{jk} = 1 \quad \text{if} \quad j = k.$$

A non-orthonormal basis with fast implementations and basis functions with finite domains is the Schoenberg basis.

The B-spline can be written as

$$g(t|\zeta_1, \dots, \zeta_K) = \sum_{k=1}^{K} \zeta_k B_k(t), \quad t \in \mathcal{T}, \tag{3.5}$$

and the least squares fits for the ζ_k are the minimizers of

$$Q(\zeta_1, \dots, \zeta_K) = \sum_{i=1}^{n} \{y_i - \sum_{k=1}^{K} \zeta_k B_k(t_i)\}^2. \tag{3.6}$$

Penalized B-Splines:

With $\zeta = (\zeta_1, \dots, \zeta_K)$, we add a penalty in (3.6) of the form $\lambda P(\zeta)$, for a penalty function

P, where λ is the penalty parameter, then proceed to minimize

$$\tilde{Q}(\zeta_1, \ldots, \zeta_K) = \sum_{i=1}^{n} \{y_i - \sum_{k=1}^{K} \zeta_k B_k(t_i)\}^2 + \lambda P(\zeta). \tag{3.7}$$

Common choices for the penalty $P(\cdot)$:

L1 penalty (Lasso penalty): $P(\zeta) = \sum_{j=1}^{K} |\zeta_j|$

L2 penalty (Ridge regression penalty): $P(\zeta) = \sum_{j=1}^{K} |\zeta_j|^2$

Roughness penalty: $P(\zeta) = \int_{\mathcal{T}} \{\sum_{k=1}^{K} \zeta_k B_k^{(2)}(t)\}^2 \, dt.$

3.3 LAB 1: GLM and Smoothing in R

3.3.1 Introduction to R

Basic Commands

- Download from CRAN
- Get Help
  - Online R Manuals
  - R for Beginners
  - R mailing lists archive for discussion about problems and solutions using R or Google
  - Command-line helps when running R

  - Type in

```
?command_name # View the help file about the usage for
                the command_name
```

  -
    ```
    help(command_name)
    ```

  -
    ```
    find(command_name) # Return the package name that a
                            command/variable belongs to
    ```

  -
    ```
    help.start() # bring up an html file of help documents
                            index page for R
    ```

- Import Data

  -
    ```
    read.table(), read.csv(), read.delim()
    ```

  -
    ```
    data()
    ```

  -
    ```
    load()
    ```

  -
    ```
    scan()
    ```
- Export Data

  -
    ```
    write.table()
    ```

  -
    ```
    save(), save.image()
    ```

Simple Multiple Linear Regression Example

- Download Longley Data via the following url:
  http://www.itl.nist.gov/div898/strd/lls/data/LINKS/DATA/Longley.dat
- Cut and paste the data into a text file, say “longley.txt” and save it.
- Then read the data into R by

  ```
  d = read.table(‘.../longley.txt’, header = TRUE)
  ```

  Here, ‘...’ should be filled with the corresponding directory where you saved the data file.

  The imported data is saved as a data.frame into the R object ‘d’. Check what ‘d’ is with

```
 class(d)
```

- Alternatively, you can also read the data into R using the url of the data file directly:

```
 d = read.table(‘http://www.itl.nist.gov/div898/strd/
lls/data/LINKS/DATA/Longley.dat’, header = TRUE)
```

- Basic commands:

```
d           # View the data
names(d)    # Check the variable names in the data
summary(d)  # Five number summary for all variables
pairs(d)    # Obtain pairwise scatter plot
```

- Run a linear regression model by using the command “*lm()*”, which is a powerful function for linear regression with continuous and categorical predictors.

  1. Fit a linear model with all predictors

     ```
     fit1 = lm(y ~ . , data = d)
     ```

     which is equivalent to

     ```
     fit1 = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6, data = d)
     ```

  2. Fit a model that includes all variables except for intercept term

     ```
     fit2 = lm(y ~ 0 + ., data = d)
     ```

     which is equivalent to

     ```
     fit2 = lm(y ~ . - 1, data = d)
     ```

  3. Fit a model that contains “x_1” and “x_3” and their interactions

     ```
     fit3 = lm(y ~ x1 + x3, data = d)
     ```

  4. Fit a model that contains “x_1”, “x_2”, “x_3” and their interactions

```
fit4 = lm(y ~ x1 * x2 * x3, data = d)
```

which is equivalent to

```
fit4 = lm(y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3
          + x1:x2:x3, data = d)
```

5. Fit a model that contains "x_1", "x_2", "x_5" and interaction between "x_1" and "x_2"

```
fit5 = lm(y ~ x1 + x2 + x5 + x1:x2, data = d)
```

- Go back to the "fit1". To see what information is contained in "fit1", use the command

```
names(fit1)
```

You will see

```
[1] "coefficients"  "residuals" "effects" "rank"
[5] "fitted.values" "assign"    "qr"      "df.residual"
[9] "xlevels"       "call"      "terms"   "models"
```

- To extract the fitted values, use the operator $.

```
fit1$fitted.values
```

- To see summary statistics from the fitted model

```
summary(fit1)
```

- ANOVA type summary

```
summary(aov(fit1))
```

- AIC based model selection for "fit1"

```
library(MASS)
stepAIC(fit1)
```

- Finally, graphical diagnostics to check the model assumptions:

  ```
  par(mfrow= c(1,2))
  plot(fit1$fitted.values, fit1$residuals, xlab = "Fitted Values",
  ylab = "Residuals", main = "Residuals Vs Fitted Values")
  abline(0,0, col = "red")
  qqnorm(fit1$residuals)
  qqline(fit1$residuals, col = "red")
  ```

- You can also save the object "fit1" to a file

  ```
  save(fit1, file = "longleyfit.Rdata") # in current working directory
  ```

  The data can be retrieved by

  ```
  load("longleyfit.Rdata")
  ```

Categorical Variables in R The lowest level is always considered to be the baseline in R (all dummy variables are 0). For example, there are level 1, 2, and 3 in a treatment, R will treat "1" as a default baseline level and will calculate the coefficients of the dummy variables corresponding to level 2 and 3. We can create two dummy variables to achieve the same result. To specify other baseline levels for the treatment, we can define dummy variables by ourselves. As an exercise, try to compare the outcome of the fitted models.

Dummy variables: Variables taking either 0 or 1, indicating if the observation has (1) the corresponding level for the treatment or not (0). For each categorical treatment, we need $\#Levels - 1$ dummy variables.

3.3.2 GLM Examples

- Poisson Regression Example

  ```
  ## Dobson (1990) Page 93: Randomized Controlled Trial :
  ```

```
counts = c(18,17,15,20,10,20,25,13,12)

outcome = gl(3,1,9) #generate a factor of 3 levels

treatment = gl(3,3) #generate a factor of 3 levels

print(d.AD = data.frame(treatment, outcome, counts))
```

```
    treatment outcome counts
 1      1         1      18
 2      1         2      17
 3      1         3      15
 4      2         1      20
 5      2         2      10
 6      2         3      20
 7      3         1      25
 8      3         2      13
 9      3         3      12
```

```
glm.D93 = glm(counts ~ outcome + treatment, family=poisson())
#model selection through deviance table
anova(glm.D93, test = "Chisq") # test can take "F" or "Cp" as well
library(MASS)
stepAIC(glm.D93) # model selection through AIC criterion
summary(glm.D93) # parameter estimates
```

Notes:

1. There are several generic functions that also work for **glm**() and they are **summary**(), **residuals**(), **anova**(), **print**(), **coef**(), **deviance**(), **AIC**(), **logLik**() and **fitted**(). To learn more about these functions, use the command

   ```
   ?command_name.glm
   ```

   For example,

```
?summary.glm
```

shows the documentation for how the summary function applies to the glm object.

2. For a Poisson regression, the responses are usually counts/integers, the underlying exponential family distribution is the **Poisson distribution**. To learn more about how to specify the link function through the variable family, type

    ```
    ?family
    ```

3. To obtain the variance-covariance matrix of the model coefficients for constructing confidence region for coefficient estimates

    ```
    ?vcov()
    vcov(glm.D93)
    ```

- Binomial Regression Example

  Specify the response as a two-column matrix:

  Smoking, Alcohol and Esophageal Cancer Example

  ```
  data(esoph)
  summary(esoph)
  # effects of alcohol, tobacco and interaction, age adjusted
  model1 = glm(cbind(ncases, ncontrols) ~ agegp + tobgp * alcgp,
  data = esoph, family = binomial())
  anova(model1)
  ```

  For more details about the data in this example

  ```
  ?esoph
  ```

3.3.3 Smoothing

Given scatterplot data (x_i, y_i), $i = 1, \ldots, n$, goal is to estimate the conditional mean $g(x) = E(y|X = x)$ via smoothing with various tools in R.

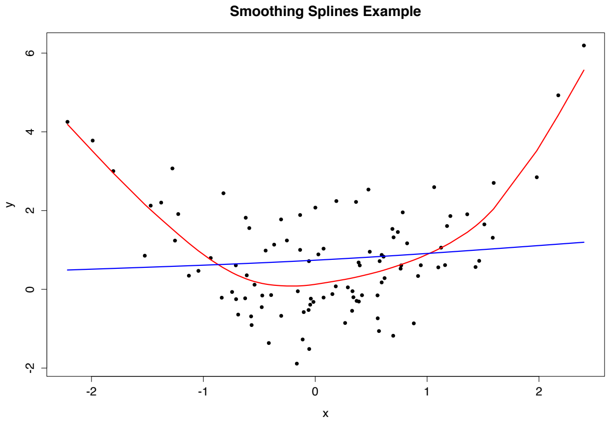
1. Cubic Smoothing Splines:

```
?smooth.spline
```

- “cv”: TRUE or FALSE (default), specifying whether to use CV or GCV (FALSE) to select bandwidth.
- “spar”: smoothing parameter, typically (but not necessarily) in $(0, 1]$. Monotone to the penalty parameter λ for the roughness of the fit (squared second derivative $\int g^{(2)}(x)^2 dx$).

Example

```
set.seed(1)
x = rnorm(100); y = x^2+rnorm(100) # simulate a quadratic relationship
fit.cv = smooth.spline(x,y, cv = TRUE) # CV fit
plot(x,y); lines(smooth.spline(x, y, cv = TRUE), col = "red")
fit.cv$spar   # the smoothing parameter corresponding to the CV choice
newspar = fit.cv$spar + 0.5
lines(smooth.spline(x,y, spar = newspar), col = "blue") # newspar fit
```



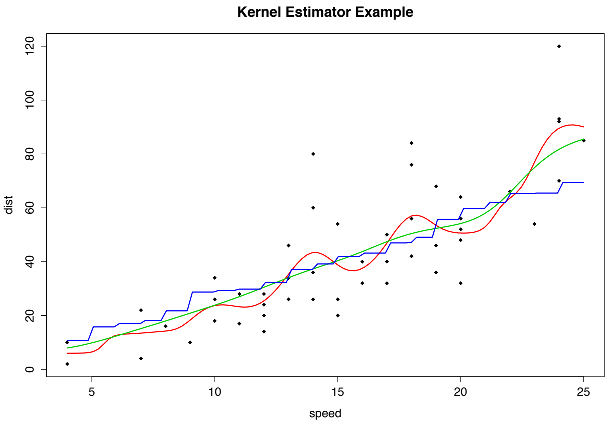
2. Kernel Estimators:

```
?ksmooth
```

- "kernel": "box" (rectangular) or "normal" (Gaussian), the kernel function to be used.
- "bandwidth": the bandwidth.

Example

```
attach(cars); plot(speed, dist, main="Kernel Estimator Example", pch=18)
lines(ksmooth(speed, dist, "normal", bandwidth=2), col="red", lwd=2)
lines(ksmooth(speed, dist, "normal", bandwidth=5), col="green", lwd=2)
lines(ksmooth(speed, dist, "box", bandwidth=10), col="blue", lwd=2)
```



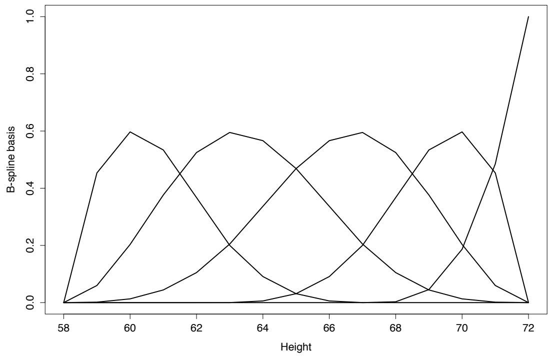
3. B-Spline

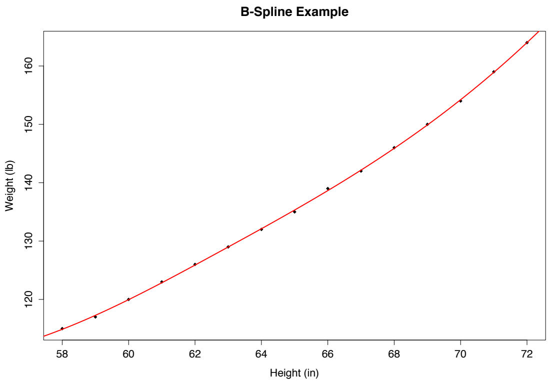
```
library(splines)
?bs
```

- "df": degrees of freedom (number of basis functions), can specify only "df" rather than "knots"
- "degree": the degree of smoothness in B-spline, 3 for cubic
- "knots": the internal breakpoints that define the spline basis
- Key relationship: $\#knots = df - degree$

Example

```
attach(women) # Women height-weight data

# Visualization of the spline basis functions

plot(58:72,bs(women$height, df = 5)[,1], type=’l’,ylim=c(0,1))

for(i in 2:5){lines(58:72,bs(women$height, df = 5)[,i])}

# Fit a linear model with B-splines

summary(fm1 <- lm(weight ~ bs(height, df = 5), data = women))

# check the prediction

plot(women, xlab = "Height (in)", ylab = "Weight (lb)",

     main="B-Spline Example", pch=18)

ht <- seq(57, 73, len = 200)

lines(ht, predict(fm1, data.frame(height=ht)), lwd=2, col="red")
```





4. Local Polynomial Smoothing

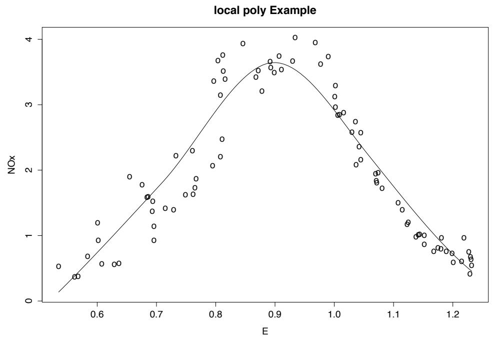
```
library(locfit)

?locfit

?lp
```

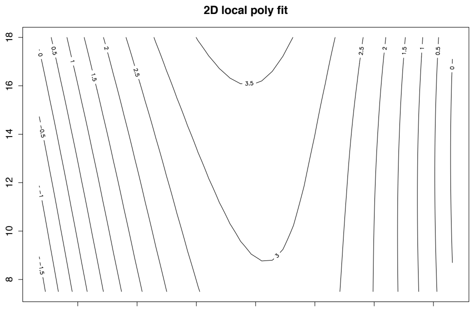
- “lp” is a local polynomial model term for locfit models, RHS of the model formula:
  - “nn”, “h” and “adpen” are 3 smoothing parameters, only need to specify one.
  - “deg”: degree of polynomials to use

- Can also do 2D smoothing

Example

```
data(ethanol, package="locfit") # ethanol gas emission data

head(ethanol)

fit1 <- locfit(NOx ~ E, data=ethanol)

summary(fit1)

plot(fit, get.data=TRUE, main="local poly Example")

# a bivariate local regression with smaller smoothing parameter

fit2 <- locfit(NOx~lp(E,C,nn=2,scale=0), data=ethanol)

plot(fit, main="2D local poly fit")
```





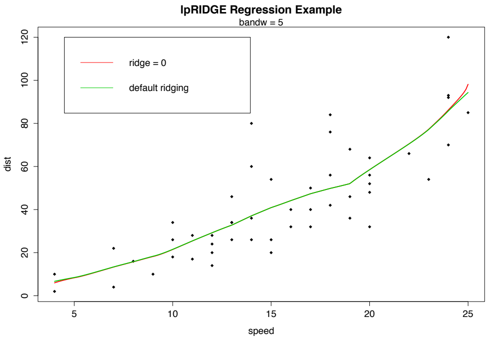
5. Local Polynomial Ridge Regression

```
library(lpridge)

?lpridge
```

- "bandwidth": smoothing parameter
- "ridge": ridging parameter governing the smoothness penalty. 0 means no penalty, default: slight ridging
- "deriv": the order of derivative to be estimated
- "order": order of polynomials used for smoothing, default: *deriv+1*

Example

```
attach(cars); plot(speed, dist, main = "lpRIDGE Regression")

myfit <- lpridge(speed,dist,bandw = 5, ridge=0) # no ridging

lines(myfit$x.out,myfit$est,col=2)

myridge <- lpridge(speed,dist,bandw = 5) # slight ridging

lines(myridge$x.out,myridge$est,col=3); mtext("bandw = 5")

legend(5,120, c("ridge = 0", "default ridging"), col = 2:3, lty = 1)
```



3.4 PROBLEM SET 1

For all problems: For each problem you need to submit a text describing your solution as well as annotated output and graphs as needed for the data analysis and computing assignments. Any computing output that you submit needs to be explicitly referred to in your text and placed in your narrative, with detailed descriptions an d explanations. Output that is not annotated or referred to will not be considered, so be very selective regarding what outputs, tables and figures you include. The default software is R. However, unless the software to be used is specified in the problem, you are free to use any software that you judge to be suitable to solve the problem.

1. Assume that vectors (X, Y) are bivariate normal with mean vector (μ_1, μ_2) and covariance matrix $(\sigma_{jk})_{1\le j,k\le 2}$. (a) Obtain $E(Y|X = x)$ for any fixed $x \in \mathcal{R}$ (b) Obtain the conditional distribution $L(Y|X = x)$ for any fixed x (c) Why would one be interested in $L(Y|X = x)$?

2. Prove (2.12).

3. Sketch how one obtains the confidence interval for θ_{0k} that is shown in example 2 on p.8.

4. Assume a researcher who feeds different diets to mice aims to establish that in a regression model with two predictors with outcome “age at death” (longevity) and states that the scientific hypothesis of interest is that “the effect of the first predictor (amount of antioxidants in the diet) on the outcome is positive and that of the second predictor (total calories in the diet) is negative.” Indicate how to proceed in a model with two predictors. Provide relevant hypotheses, describe limit distributions and how to obtain a suitable p-value, based on (2.12). How do you modify your answer if the researcher wants to establish that the two effects have different signs, i.e., one is positive and the other one is negative, irrespective of the sign of the effect of the first predictor?

5. Show that for any given kernel function K and bandwidth h, the kernel estimator of local linear type when setting $\beta_1 = 0$ in the weighted least squares expression that is to be minimized, i.e., when fitting local constants by weighted least squares, is identical to the quotient type kernel estimator.

6. The data set melanoma.dat that you can download from the class web site includes observed number of melanomas per one million people in Connecticut, the year in which the melanomas have been diagnosed, and the number of sunspots during that year as a measure of solar activity. (Please read up on biomedical terms unknown to you throughout the class). Melanomas have a complex etiology and are thought to be caused by several factors, including heredity, hormonal influences and sun exposure. The data set was generated to study the relationship between solar activity, calendar year and melanoma incidence. (a) Analyze the data with a Poisson regression model. Do the predictors have an effect on the response? (b) Which effects are significant? Write all hypotheses you consider. (c) Write hypotheses and find the p-value for one-sided tests, where we are only interested in whether the number of melanomas increases as the number of sunspots increases and whether their number increases over the calendar years, i.e., whether more recently melanomas are more frequently encountered than earlier.

7. For the two slope parameters in the previous problem, assuming that the estimators you obtain by fitting the GLM satisfy a result like (2.10) for the covariance matrix Σ of the parameter estimates, construct the simultaneous 95% confidence ellipsoid for the two slope parameters. Use the simultaneous confidence ellipsoid construction to determine a $p-$ value for the null hypothesis that both slope parameters are 0.

8. Test the null hypothesis that the slope parameter for year is smaller or equal to that for solar activity. Write null hypothesis and alternative and provide the details for your test.

9. Assume in an observational study we would like to analyze the relationship of systolic

blood pressure (SBP) (Y) and body mass index (BMI) (X) for a sample of n randomly selected subjects. The data are recorded as (X_i, Y_i), $i = 1, \dots, n$.

In a first exploratory analysis, you would like to study the relationship by a smoothing approach. Let $\hat{m}(x)$ be a reasonable nonparametric smoothing estimate for $m(x) = E(Y|X = x)$, for all x in the domain of m. (a) Give one explicit example of a smoothing estimate $\hat{m}$ which is (asymptotically) consistent, i.e., converges in probability to $m(x)$ as the sample size increases (no proof required). (b) Identify the smoothing parameter. In which way do bias and variance depend on its choice? (c) Using any consistent estimator $\hat{m}(x)$ of $m(x)$, obtain an estimator $\hat{v}(x)$ for the function $v(x) = \text{var}(Y|X = x)$, using the data (X_i, Y_i), $i = 1, \dots, n$. (d) Is the estimator $\hat{v}(x)$ also consistent?

10. Provide examples of pdfs $f(x, \theta)$ where one does not have exchangeability of integral and differentiation.

11. Derive the third Bartlett identity (2.4). Indicate the result you need re. the dominating function (you are not asked to find such a function).

12. Derive first and second Bartlett identity for the $p-$dimensional case.

13. The data set zmk.dat consists of a predictor which is a voltage measured in a chewing muscle and a force which is exercised by the muscle. The data is in the form (voltage, force), with voltage the only predictor variable. Smooth the scatterplot (X_i, Y_i) to obtain an estimate of the mean regression function $\hat{\mu}(x) = \hat{E}(Y|X = x)$. Find a suitable program for smoothing, where you can control the choice of the smoothing parameter. Criteria for comparing curve fits are to avoid *undersmoothing* (too many small little bumps in the curves) and *oversmoothing* (fitted curves have too many consecutive data points on one side of the curve, a sign of lack-of-fit). (a) Obtain two curves which are underfitted and two which are overfitted (b) Find a subjectively "good" smooth fit (d) Describe your strategy to obtain a good fit.

14. Consider p-dimensional predictor vectors x_i, $i = 1, \dots, n$, that we assume to be equidistantly spaced (to the extent possible) on the unit cube $[0, 1]^p$ and for which we have responses $Y_i = g(x_i) + \epsilon_i$, where the errors ϵ_i are i.i.d. with mean zero and variance $\sigma^2 < \infty$. As estimator for the regression function $g(x)$ we place a ball $B_h(x)$ with radius h around x as midpoint and take the average of the Y_i for which the corresponding x_i fall into the ball, i.e. $\hat{g}(x) = \frac{1}{\sum 1_{\{x_i \in B_h(x)\}}} \sum Y_i 1_{\{x_i \in B_h(x)\}}$. We assume $h = h_n \to 0$ as $n \to \infty$.

    Show that (a) $\text{var}(\hat{g}(x)) = O(\frac{1}{nh^p})$

    (b) $E(\hat{g}(x)) - g(x) = O(h^2)$

    (c) From (a), (b) obtain the rate of the MSE of $\hat{g}(x)$.

    (d) Setting $h = n^{-\alpha}$ find the α such that the MSE is minimal.

(e) Plugging in this value of h, obtain the rate of convergence of the MSE.

(f) Discuss the curse of dimensionality that can be deduced from this rate.

15. Assume a researcher studies the effects of 3 predictors on an outcome and aims to show that the effect of the first predictor is larger than the effects of the other two predictors combined. Assuming that "effect" means "absolute value of the slope parameter", indicate how to proceed. Provide relevant hypotheses, describe limit distributions and describe how to obtain a suitable p-value.

16. Sketch the derivation of (2.6) by using the CLT.

17. Derive (2.7) from (2.6) (Hint: Taylor expansion; you do not have to find exact bounds for the remainder terms).

18. Derive (2.8) from (2.7).

19. Obtain explicit formulas for the minimizers (in terms of the ξ_k) in (3.7) for the roughness penalty.

20. Write a general linear function estimator derived from a scatterplot (X_i, Y_i) as $\hat{g}(x) = \sum_{i=1}^{n} w_i(x,h)Y_i$, where h is a bandwidth. Show that the local linear estimator can be represented as a general linear function estimator and find the weights $w(x,h)$.

21. Prove (2.13).

22. Obtain explicit formulas for the minimizer $\hat{\beta}_0(x)$ of local linear estimators (3.4), for a kernel K which is a pdf with domain $[-1,1]$.

4 The Generalized Linear Model (GLM)

4.1 Introduction

Data (x_i, y_i), $i = 1, \cdots, n$, $\quad Ey = \mu$

Multiple Linear Regression Model

1. $\mu = \sum_{j=0}^{p-1} x_j\beta_j$, where $x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{p-1} \end{pmatrix}$, $x_0 = 1$, is the vector of predictors, and

$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$ is the parameter vector.

Shorthand:

$$\mu_{n\times 1} = X_{n\times p}\beta_{p\times 1},$$

where X is the *design matrix* and β is the parameter vector. (”systematic part”).

2. Errors ϵ_i independent, normal, with zero mean and constant variance (homoscedastic errors), i.e., $E\epsilon_i = 0$, $\text{var}(\epsilon_i)^2 = \sigma^2$, adopting the fixed design regression model,

$$y = X\beta + \epsilon$$

More general view:

- *Random component*: y_i independent, $Ey_i = \mu_i$, y_i normal with constant variance, i.e., $y_i \sim N(\mu_i, \sigma^2)$.
- *Systematic component*: Linear predictor η defined by $\eta = X\beta$.
- *Link function*: $\mu = \eta$ (”identity link”)

Extending beyond the classical regression model: generalize normality assumption in random part and the link function in the systematic part.
Generalizing to: Exponential family distribution for the random part, monotone differentiable link function for the systematic part.

4.2 Components of the GLM

- Systematic component:
  Linear predictor $\eta = X\beta$ (G1)
  Link function $\eta = g(\mu), \quad \mu = Ey$ (G2)
  (this is the notation we use, while some other authors use the notation $\mu = g(\eta)$)

- Random component:

  $Ey = \mu, \quad y \sim \textit{exponential family.}$ (G3)

4.3 Exponential Family

A r.v. y has a distribution in the *exponential family*, if the pdf or pmf of y has the form

$$f(y, \theta, \phi) = \exp\{(y\theta - b(\theta))/\phi + c(y, \phi)\} \tag{4.1}$$

for some functions $b(\cdot)$, $c(\cdot, \cdot)$, where we usually require $\phi > 0$. This is the *canonical form* with canonical parameter θ, and is sometimes referred to as canonical dispersion family. Covers cases $y \in \mathcal{R}^1$, $\theta \in \mathcal{R}^1$ as well as $y, \theta \in \mathcal{R}^p$; if $p > 1$, replace $y\theta$ by $y^\top \theta$, $b : \mathcal{R}^p \to \mathcal{R}$; usually $\phi \in \mathcal{R}$.

Example: Normal distribution

$$\begin{aligned} f(y, \theta, \phi) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2} \\ &= \exp\{(y\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2))\} \end{aligned}$$

Setting

$$\theta = \mu, \;\; \phi = \sigma^2, \;\; b(\theta) = \frac{\theta^2}{2}, \;\; c(y, \phi) = -\frac{1}{2}(\frac{y^2}{\phi} + \log(2\pi\phi)),$$

this is in the exponential family with canonical parameter μ.

Also the Gamma distribution is in the exponential family. This is easiest seen by writing the Gamma density with parameters $\mu, \nu > 0$ and pdf

$$f_{\mu,\nu}(y) = \frac{1}{\Gamma(\nu)} (\frac{\nu y}{\mu})^\nu \exp(-\frac{\nu y}{\mu}) \frac{1}{y}, \;\; y \geq 0.$$

Note the relation with the usual parametrization of the Gamma density which is (for $\kappa, \lambda > 0$)

$$f_{\kappa,\lambda}(y) = \lambda^\kappa y^{\kappa-1} e^{-\lambda y}/\Gamma(\kappa), \;\; y \geq 0.$$

Cumulant Generating Function

For a r.v. Y with a given distribution F and pdf f define the moment generating function $M(t)$

$$M(t) = \int \exp(ty) f(y)\, dy = E(\exp(tY)) = E \sum_{k=0}^{\infty} \frac{(tY)^k}{k!} \tag{4.2}$$

and the cumulant generating function $C(t)$ as

$$C(t) = \log(M(t)) = \sum_{j=0}^{\infty} \kappa_j \frac{t^j}{j!}. \tag{4.3}$$

It holds

$$\frac{\partial^q}{\partial t^q} M(t)|_{t=0} = E(Y^q) = \mu_q, \quad \frac{\partial^q}{\partial t^q} C(t)|_{t=0} = \kappa_q, \quad q \geq 1,$$

where κ_q is the q-th cumulant of the distribution of Y. It can be shown that

$$\kappa_0 = 0, \quad \kappa_1 = EY, \quad \kappa_j = E(Y - EY)^j, \quad j = 2, 3. \tag{4.4}$$

Plugging into (4.2) the pdf of an exponential family distribution,

$$M(t) = \int \exp\left[\frac{1}{\phi}\{y(\theta + t\phi) - b(\theta) + b(\theta + t\phi) - b(\theta + t\phi)\} + c(y, \phi)\right] dy = \exp\left(\frac{b(\theta + t\phi) - b(\theta)}{\phi}\right),$$

whence

$$C(t) = [b(\theta + t\phi) - b(\theta)]/\phi.$$

This implies that $b(\theta)$ generates the cumulants as follows:

$$\kappa_q = \frac{\partial^q}{\partial t^q} C(t)|_{t=0} = \phi^{q-1} b^{(q)}(\theta). \tag{4.5}$$

Example: Gaussian distribution. Here $b(\theta) = \frac{\theta^2}{2}$, $\phi = \sigma^2$ and

$$\begin{aligned}
EY &= \mu_1 = \kappa_1 = b^{(1)}(\theta) = \theta \\
\mathrm{var}(Y) &= E(Y - EY)^2 = \kappa_2 = \sigma^2 b^{(2)}(\theta) = \sigma^2 \\
E(Y - EY)^3 &= \kappa_3 = \sigma^4 b^{(3)}(\theta) = 0
\end{aligned}$$

4.4 Likelihood in the exponential family

Consider $\theta \in \mathcal{R}^1$. For the log likelihood,

$$l(\theta, \phi, y) = \log f(y, \theta, \phi) = (y\theta - b(\theta))/\phi + c(y, \phi)$$

$$\begin{aligned}
\frac{\partial l}{\partial \theta} &= (y - b'(\theta))/\phi \\
\frac{\partial^2 l}{\partial \theta^2} &= -b''(\theta)/\phi.
\end{aligned}$$

From the first Bartlett identity (2.2),

$$0 = E_\theta(\frac{\partial l}{\partial \theta}) = (\mu - b'(\theta))/\phi$$

therefore

$$E_\theta(y) = \mu = b'(\theta). \tag{4.6}$$

From the second Bartlett identity (2.3),

$$\begin{aligned}
0 &= E_\theta(\frac{\partial^2 l}{\partial \theta^2}) + E_\theta(\frac{\partial l}{\partial \theta})^2 \\
&= -\frac{b''(\theta)}{\phi} + E_\theta(\frac{y - b'(\theta)}{\phi})^2 \\
&= -\frac{b''(\theta)}{\phi} + \frac{\mathrm{var}_\theta(y)}{\phi^2},
\end{aligned}$$

by (4.3), therefore

$$\mathrm{var}_\theta(y) = b''(\theta)\phi. \tag{4.7}$$

4.5 Variance functions

Note: $V_0(\theta) = b''(\theta)$ is the part of $\mathrm{var}_\theta(y)$ that depends on θ in (4.7).

Transforming into a function of μ: By (4.6), $\mu = \mu(\theta) = b'(\theta)$.

Assume $\mathrm{var}_\theta(y) > 0$, then $b''(\theta) > 0$, so that the function

$$b' : \theta \to \mu \qquad \text{is invertible,}$$

$$b'^{-1} : \mu \to \theta, \quad \theta = b'^{-1}(\mu) = \theta(\mu)$$

Then we can define the variance function V as a function of the argument μ,

$$V(\mu) = V_0(\theta(\mu)) = b''(\theta(\mu)) = b''(b'^{-1}(\mu))$$

Example: Gaussian model:

$$b(\theta) = \frac{\theta^2}{2}, \;\; b'(\theta) = \theta, \;\; b''(\theta) = 1 > 0,$$

$$V(\mu) = b''(b'^{-1}(\mu)) = b''(\mu) = 1,$$

$\mu = b'(\theta) = \theta;$ this is a situation with constant variance function.

5 Poisson and Binomial Regression

5.1 Poisson family

L.von Bortkewitsch (1898): "The Law of Small Numbers". An early historical example where the number of deaths from horse kicks in the Prussian army 1875-1894 is analyzed.

These are *count data*. Are they Poisson? This depends critically on the relation between dispersion and mean.

Pmf for Poisson data:

$$f(y, \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = \exp\{y \log \lambda - \lambda - \log y!\}$$

$$\theta = \log \lambda, \ \ \phi = 1, \ \ b(\theta) = \lambda = \exp(\theta), \ \ c(y, \phi) = -\log y!$$

$$Ey = \mu = b'(\theta) = \exp(\theta) = \lambda$$

$$\mathrm{var}(y) = b''(\theta)\phi = \exp(\theta) \cdot 1 = \lambda$$

Variance function:

$$V(\mu) = b''(b'^{-1}(\mu)) = \mu,$$

since here

$$b'^{-1}(\mu) = \log(\mu) = \theta.$$

In the Poisson case, $\mathrm{var}(y) = E(y) = \mu, \ V(\mu) = \mu$. This is a central property of Poisson data.

5.2 Binomial family

Counting discrete fractions, number of successes in n trials or counting the number of lymphocytes among 100 leukocytes under the microscope.

Data: $y = \frac{count}{100}$ (a fraction), $y \sim \frac{1}{n} Bin(n, \pi)$,

$ny \sim Bin(n, \pi)$ with pmf

$$\begin{aligned} f(y, \pi, n) &= \binom{n}{ny} \pi^{ny} (1-\pi)^{n-ny} \\ &= \exp\{ny \log \pi - ny \log(1-\pi) + n \log(1-\pi) + \log \binom{n}{ny}\} \\ &= \exp\{[y(\log \frac{\pi}{1-\pi}) - (-\log(1-\pi))]/(1/n) + \log \binom{n}{ny}\} \end{aligned}$$

$$\theta = \log \frac{\pi}{1-\pi} =: \text{logit}(\pi), \quad \phi = \frac{1}{n},$$

$$\frac{\pi}{1-\pi} = e^\theta, \quad \pi(1+e^\theta) = e^\theta,$$

$$\pi = \frac{e^\theta}{1+e^\theta} =: \text{expit}(\theta).$$

Note

$$1 - \pi = \frac{1}{1+e^\theta} \tag{5.1}$$

and

$$b(\theta) = -\log(1-\pi) = -\log(\frac{1}{1+e^\theta}) = \log(1+e^\theta) \tag{5.2}$$

$$c(y, \phi) = \log \binom{1/\phi}{y/\phi}.$$

Moments:

$$Ey = \mu(\theta) = b'(\theta) = \frac{e^\theta}{1+e^\theta} = \text{expit}(\theta) = \pi,$$

$$\text{var}(y) = b''(\theta)\phi = \frac{1}{n} \; \frac{e^\theta(1+e^\theta) - e^\theta e^\theta}{(1+e^\theta)^2} = \frac{1}{n} \quad \frac{e^\theta}{(1+e^\theta)^2} = \frac{1}{n} \; \pi(1-\pi), \tag{5.3}$$

since

$$\pi = e^{\theta}/(1+e^{\theta}),\ \ 1-\pi = \frac{1}{1+e^{\theta}}.$$

Variance function:

$$b'^{-1}(\mu) = \mathrm{logit}(\mu) = \theta(\mu),$$

and using

$$e^{\theta(\mu)} = \frac{\mu}{1-\mu},$$

$$V(\mu) = b''(b'^{-1}(\mu)) = \frac{e^{\theta(\mu)}}{(1+e^{\theta(\mu)})^2} = \frac{\mu/(1-\mu)}{1/(1-\mu)^2} = \mu(1-\mu).$$

6 Link Functions

Relates linear predictor $\eta = \sum_{j=1}^{p} x_j^{\top}\beta_j = X\beta$ with $\mu = Ey$

Linear model: $\eta = \mu$

Writing $\eta = g(\mu)$, this means $g \equiv \mathrm{id}$ (identity function) for the linear model. In most applications, the identity link is not suitable, e.g., for count data, where μ is constrained to be positive while η is not. We consider only smooth and invertible link functions.

Common link functions (note the range constraints on μ as compared to η):

- log link: $\eta = \log(\mu),\quad \mu = e^{\eta},\quad \mu > 0$
- logit link: $\eta = \log(\frac{\mu}{1-\mu}) = \mathrm{logit}(\mu),\ \ \mu = \mathrm{expit}(\eta) = \frac{e^{\eta}}{1+e^{\eta}},\quad 0 < \mu < 1$
- probit link: $\eta = \Phi^{-1}(\mu),\ \ \mu = \Phi(\eta)$, $0 < \mu < 1$.
- log-log(complementary log) link: $\eta = \log(-\log(1-\mu)),\ \ \mu = 1-\exp(-\exp(\eta))$, $0 < \mu < 1$.
- power family link: $\eta = (\mu^{\lambda}-1)/\lambda,\ \ \eta = \log\mu$ for $\lambda = 0$

In the exponential family, y is *sufficient statistic* for θ (this is a consequence of the *factorization theorem*). Note a sufficient statistic $T(X)$ for a parameter θ is characterized by the fact that the conditional distribution of X given $T(X)$ is independent of θ. The factorization theorem says that if and only if the density for X has the form $f_\theta(x) = g(T(x), \theta)h(x)$ for functions g, h, T is a sufficient statistic for θ.

What is a sufficient statistic for the regression parameter β?

We address this for the case where the link g is such that $\eta = \theta$. Since

$$b'^{-1}(g^{-1}(\eta)) = b'^{-1}(\mu) = \theta,$$

$\eta = \theta$ implies that

$$b'^{-1} \circ g^{-1} \equiv id.$$

This is equivalent to

$$g(\cdot) \equiv b'^{-1}(\cdot).$$

For the pdf of the entire sample of n subjects,

$$f(\cdot) = \prod_{i=1}^{n} f(y_i, \theta_i, \phi) = \exp\{\sum_{i=1}^{n} [y_i\theta_i - b(\theta_i)]/\phi + \sum_{i=1}^{n} c(y_i, \phi)\}$$

$$\text{and observing} \quad \begin{aligned} \sum_{i=1}^{n} y_i\theta_i &= \sum_{i=1}^{n} y_i(\sum_{j=1}^{p} x_{ij}\beta_j) \qquad (\text{as } \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \theta_i) \\ &= \left(\sum_{i=1}^{n} x_{i1}y_i, \cdots, \sum_{i=1}^{n} x_{ip}y_i\right) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \end{aligned}$$

the factorization theorem implies that $x^\top y$ is the sufficient statistic for β.

The assumption $\eta = \theta$ for the link function is crucial for this to happen and therefore is called the *canonical link*. The canonical link is often chosen as default link function; it may or may not be adequate in practice.

Examples for canonical links:

normal $\theta = \mu, \;\; \eta = \theta \;\; \Rightarrow \;\; \eta = \mu$, identity link

Poisson $\theta = \log \lambda = \log \mu, \;\; \eta = \theta \;\; \Rightarrow \;\; \eta = \log \mu$, log link

Binomial $\theta = \log \frac{\pi}{1-\pi} = \log \frac{\mu}{1-\mu}, \;\; \eta = \theta \;\; \Rightarrow \eta = \text{logit}(\mu)$, logit link

Gamma $\theta = -\mu^{-1}, \;\; \eta = \theta \;\; \Rightarrow \;\; \mu = -\frac{1}{\eta}, \;\; \eta = -\frac{1}{\mu}$, inverse link

Inverse Gaussian $\theta = -\frac{1}{2\mu^2}$ (up to a factor), $\theta = \eta$ then leads to $\eta = -\frac{1}{2\mu^2}$,

where the pdf of the inverse Gaussian is

$$f(\mu, y, \sigma) = (\frac{1}{2\pi\sigma^2 y^3})^{1/2} \exp(-\frac{(y-\mu)^2}{2\sigma^2\mu^2 y}), \;\; y \geq 0.$$

7 Goodness-of-fit

7.1 Pearson type criteria

Residuals are obtained by $r_i = y_i - \hat{y}_i$, where $\hat{y}_i = \hat{\mu}_i = g^{-1}(\hat{\eta}_i) = g^{-1}(X_i\hat{\beta})$.

Pearson or standardized residuals:

$$r_{iP} = \frac{y_i - \hat{y}_i}{\{V(\hat{\mu}_i)\phi\}^{1/2}} = \frac{r_i}{st.\hat{d}ev(y_i)}$$

Example: Binomial regression

$$r_{iP} = \frac{y_i - \hat{\pi}_i}{\sqrt{n_i^{-1}\hat{\pi}_i(1-\hat{\pi}_i)}}, \quad \hat{\pi}_i = g^{-1}(\hat{\eta}_i)$$

where y_i = fraction of binomial counts out of n, as before.

Note: *Pearson's* χ^2, a goodness-of-fit statistic, is defined by

$$P = \sum_{i=1}^{n} r_{iP}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{\text{var}}(y_i)}$$

Can show: Under certain design conditions (especially relevant for discrete distributions), if the model fits the data,

$$P \sim \chi^2_{n-p},$$

where $\beta \in \mathcal{R}^p$; especially

$$E(P) \approx n - p.$$

7.2 Deviance based criteria

We defined earlier

$$D^*(y, \hat{\mu}) = 2\{l(y, y) - l(\hat{\mu}, y)\},$$

based on the comparison between saturated (each data point has its own parameters) and assumed model. In the exponential family, assuming ϕ is known,

$$D^*(y, \hat{\mu}) = 2\sum_{i=1}^{n}\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}/\phi$$

where

$$\tilde{\theta}_i = \theta(y_i) = b'^{-1}(y_i), \ \ \hat{\theta}_i = \theta(\hat{\mu}_i) = b'^{-1}(\hat{\mu}_i)$$

are the MLEs of the corresponding canonical parameters.

Example: Poisson model

$$\hat{\theta}_i = \log(\hat{\lambda}_i), \ \ \tilde{\theta}_i = \log(y_i), \ \ b(\theta_i) = \exp{(\theta_i)}$$

where we need to assume $y_i > 0$ or make a modification by adding a small constant to avoid problems when $y_i = 0$. Then

$$b(\hat{\theta}_i) = \hat{\lambda}_i, \ \ b(\tilde{\theta}_i) = y_i, \ \ \phi = 1, \quad \text{as} \quad b(\theta) = \exp(\theta), \quad \text{and}$$

$$D^*(y, \hat{\mu}) = 2\sum_{i=1}^{n}\{y_i \log\frac{y_i}{\hat{\lambda}_i} + (\hat{\lambda}_i - y_i)\}.$$

If the fitted model is the correct model, under regularity assumptions on the design, one has

$$D^*(y,\hat{\mu}) \sim \chi^2_{n-p} \quad \text{for large } n, \text{ and}$$

$$ED^*(y,\hat{\mu}) = n-p.$$

This is the same behavior as for Pearson's χ^2. If the assumed model fits the data, we expect that Pearson's criterion and deviance do not differ too much,

$$P \approx D^*(y,\hat{\mu}).$$

7.3 Deviance residuals

Write $D^* = \sum_{i=1}^{n} d_i$. Define

$$r_{Di} = \text{sign}(y_i - \hat{\mu}_i)|d_i|^{1/2},$$

the square root is used as $|d_i|$ corresponds to r_{iP}^2, necessitating the sign adjustment.

Example: Poisson

$$r_{Di} = \sqrt{2}\,\text{sign}(y_i - \hat{\lambda}_i)|y_i \log \frac{y_i}{\hat{\lambda}_i} - (y_i - \hat{\lambda}_i)|^{1/2}$$

7.4 Deviance Table

Advantage of Deviance over Pearson's χ^2:

It corresponds to log likelihood therefore is additive; can be used to compare the fit of a sequence of nested models, useful for covariate selection; the likelihood ratio test that is obtained from differences of deviances between smaller and larger models and is based on the χ^2 distribution serves the analogous function to the F-test that is used in the normal model (Wilk's Theorem).

Deviance table ("ANOVA table" and "ANOVA test") :

Sequence of models	Deviance (e.g.)	Diff	df	χ^2_p $(p=1)$
X_1, X_2, X_3	50			
X_1, X_2	70	20	1	$P < 10^{-3}$
X_1	100	30	1	.
0	200	100	1	.

In this example we clearly need all 3 predictors. A general test is based on

$\log(LR) = D(\text{smaller model}) - D(\text{larger model}) \sim \chi^2_p, \quad p = df(\text{larger model}) - df(\text{smaller model}),$

where D is the deviance of the corresponding model.

An important application of this LR test is testing $H_0: \ \beta_2 = \ldots = \beta_p = 0$, the test that all slopes vanish. This is the test for overall regression effect. If we can reject the null hypothesis, and thus establish a regression effect, we can continue to develop a meaningful regression analysis.

Principles of *model selection* apply here, for example stepwise forward or backward variable selection as known from multiple linear regression, where the F-statistic is replaced by deviance, or the *AIC (BIC) criterion*, where one aims to minimize

$$AIC(p) = D(p) + 2p, \quad BIC(p) = D(p) + p \log n,$$

respectively, where $D(p)$ is the deviance of a model that includes p predictors and the second term is the penalty for the number of included parameters.

These selection criteria work for small to moderately large numbers of predictors, but not for large numbers. In the latter case, methods that use a $L1$ type penalty, such as Lasso, Elastic net etc., generally are preferred.

7.5 LAB 2: Model Selection and Diagnostics for GLM

7.5.1 Model Selection in GLM

- **AIC Type Criterion**

  In library “MASS”,

  ```
  library(MASS)
  stepAIC(glmfit)
  ```

  allows you to choose the best model starting from the input model “glmfit” through AIC type criteria.

  Options include:

  - “direction”: model selection direction, can be either of “forward”, “backward”, and with default “both”.
  - “scope”: a list of upper and lower scopes of models to be considered. Both models are specified as model formulae (see Example below). If no scope is specified, the default upper is the model in “glmfit” and “backward” selection is used.
  - “k”: the constant weight for measuring the model complexity. Default: $k = 2$ (AIC); set $k = \log(n)$ to use BIC.
  - “trace”: if positive, print out the selection process.

  Example

  ```
  library(MASS)
  example(birthwt) # birth weight example
  birthwt.glm <- glm(low~age+lwt+race+smoke+ptd,
                     family = binomial(), data = bwt)
  birthwt.step <- stepAIC(birthwt.glm, trace = FALSE)
  birthwt.step$anova
  # define a wider scope with more predictors in the upper model
  ```

```
Scope = list(upper = ~age+lwt+race+smoke+ptd+ht+ui+ftv, lower = ~1)
birthwt.step2 <- stepAIC(birthwt.glm, trace = FALSE, scope = Scope)
birthwt.step2$anova
```

- **Deviance Table**

  In R, use the following command to obtain the deviance table.

  ```
  anova(glmfit, test = "Chi")
  ```

  Note:

  1. It is a sequential variable selection method, and is sensitive to the order of parameters.
  2. In the model specification, the model selection is starting with an intercept only model, and sequentially testing additional terms to enter the model based on likelihood ratio tests.

  Example

  ```
  # use the previous birth weight example
  anova(glm(low~smoke+ptd+lwt, family = binomial(), data = bwt), test="Chi")
        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
  NULL                    188      234.67
  smoke  1   4.8674       187      229.81  0.02737 *
  ptd    1  10.4757       186      219.33  0.00121 **
  lwt    1   4.1060       185      215.22  0.04273 *

  anova(glm(low~lwt+ptd+smoke, family = binomial(), data = bwt), test="Chi")
        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
  NULL                    188      234.67
  lwt    1   5.9813       187      228.69 0.0144581 *
  ptd    1  11.1939       186      217.50 0.0008207 ***
  smoke  1   2.2739       185      215.22 0.1315675
  ```

- **Likelihood Ratio Test**

  If one likes to investigate whether a reduced model is preferred, a likelihood ratio test can be applied for GLM, which is performed by "anova()". Note that this test can only be used for nested models, where the null model is the smaller model (a special case of the larger model), and the alternative is the larger model, although the function "anova()" does not need to specify the order of the two models.

  Example

```
bwtfit <- glm(formula = low ~ lwt + race + smoke + ptd, family = binomial(),
    data = bwt)
h0.fit = glm(low~lwt + race + smoke, family=binomial(),
                   data = bwt)
anova(h0.fit, bwtfit, test="Chi")


Model 1: low ~ lwt + race + smoke
Model 2: low ~ lwt + race + smoke + ptd
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       184     215.01
2       183     207.04  1   7.9752 0.004742 **
```

  In this example, we are basically testing the null hypothesis that the slope parameter for "ptd" is 0, and the small p-value indicates that the null model is rejected. Therefore, the larger model is more appropriate.

7.5.2 Model Diagnostics

We illustrate tools for model diagnostics using the previous birth weight data example. We start with the logistic regression model selected based on AIC from the previous example:

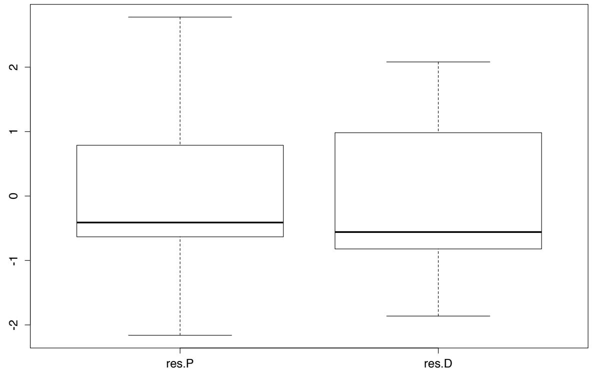
```
bwtfit <- glm(formula = low ~ lwt + race + smoke + ptd, family = binomial(),
    data = bwt)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.380164   0.915144  -0.415   0.6778
lwt         -0.012046   0.006463  -1.864   0.0623 .
raceblack    1.278212   0.520349   2.456   0.0140 *
raceother    0.898946   0.423585   2.122   0.0338 *
smokeTRUE    0.877065   0.391881   2.238   0.0252 *
ptdTRUE      1.223296   0.436828   2.800   0.0051 **
```

- Deviance Residuals and Pearson Residuals

  If the two kinds of residuals are not quite similar to each other, the model may suffer from potential lack-of-fit. You can retrieve both kinds of residuals by using

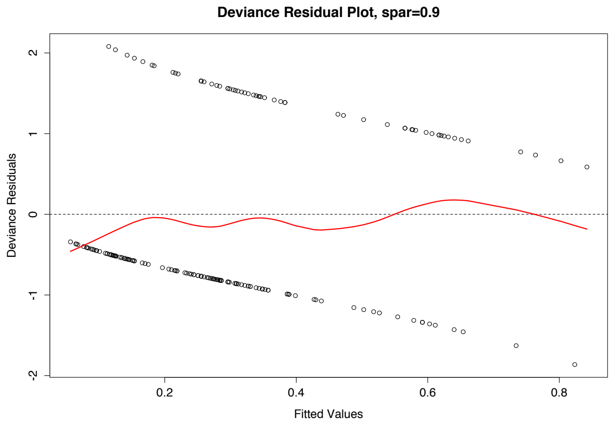
  ```
  res.P = residuals(bwtfit, type="pearson")
  res.D = residuals(bwtfit, type="deviance") #or residuals(fit), by default
  boxplot(cbind(res.P, res.D),  names = c("Pearson", "Deviance"))
  ```

  The boxplot shows similar distributions of the two types of residuals, so this particular diagnostic does not provide any indication for lack-of-fit.

- Residual Plots

  Plot residuals against fitted values, where the latter can be obtained by “bwtfit$fitted.values”.



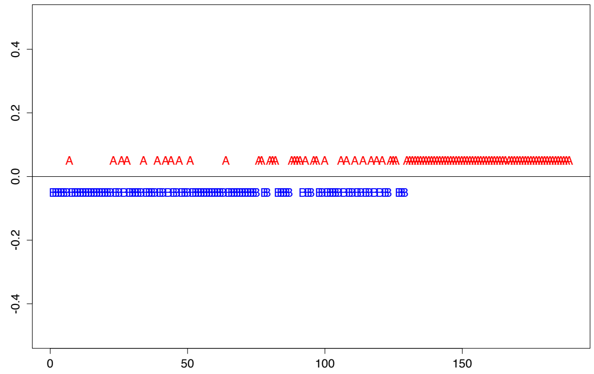
The aim is to check if there are any systematic patterns left in the residuals.

The scatter plot itself does not provide much information because of the special type of binary response type in logistic regression (why?). To aid our decision-making process, we turn to smoothers for help. It is useful to complement the deviance residual plot with an overlaying smoothing splines fit. In the above figure the red curve is the splines fit, which is quite close to 0, but may have a slight quadratic pattern. This might indicate a slight lack-of-fit of the model, and higher order terms or interaction terms can be added to see if the pattern persists.

- Runs Test

  Another handy tool is the runs test, which tests against the null hypothesis that there are no systematic patterns in a sequence of random numbers, here the residuals.

```
library(lawstat) # please pay attention to the library
# there are different runs.test() functions in different packages
# we are specifically using this one!
runs.test(y = res.D, plot.it = TRUE)

         Runs Test - Two sided

data:  res.D
Standardized Runs Statistic = -6.345, p-value = 2.224e-10
```



We always do the two-sided test, because any systematic trend (no matter positive/negative correlated) suffices to reject the goodness-of-fit of the model. Here the output p-value is very small, because a number of consecutive positive (negative) residuals are on the right (left), corresponding to large (small) fitted values, according to the following plot. Therefore, the test indicates lack-of-fit in the model.

Note: the function with the same name in another package assumes Gaussian distribution, which is violated for GLM residuals. Thus we only use this package, which assumes no distribution for the input vector.

- Leverage Points

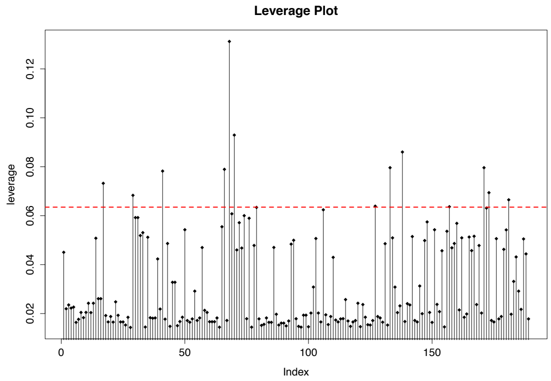
  To identify influential data points, plot the leverage h_{ii} (diagonal of hat matrix) against the index of the points. An observation is suspected as a leverage point if $h_{ii} > 2p/n$, where p is the number of coefficients in the model and n sample size, respectively.

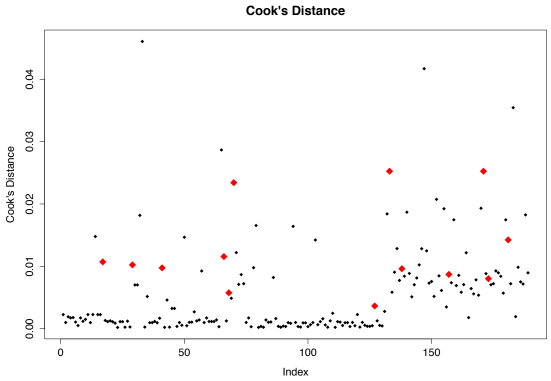
```
leverage = hatvalues(bwtfit)
plot(names(leverage), leverage, xlab="Index", type="h")
points(names(leverage), leverage)
abline(h=2*length(bwtfit$coefficients)/nrow(birthwt),col=2,lwd=2,lty=2)
```

- Cook’s Distance

  To detect outliers/influential observations, you can use Cook’s distance obtained by

```
cooks = cooks.distance(bwtfit)
plot(cooks, main="Cook's Distance", ylab="Cook's Distance")
```





Note: the relationship between outliers and influential observations is not trivial. Outlier does not necessarily imply leverage point and vice versa (why?). We can first identify influential points based on the leverage plot, and then check their Cook’s distance. Carefully checking for extreme values in all predictors can be tedious, but is the best way to justify that they are true outliers.

7.5.3 Alternative Formulation of Logistic Regression in R

Suppose we have a logistic regression data with binary group response.

- Response: number of successes called “Yes”, and number of failures called “No” as column 1 and 2
- Predictor 1: categorical variable called “A” with levels 1, 2, 3 as column 3
- Predictor 2: categorical variable called “B” with levels 1, 2 as column 4
- Predictor 3: categorical variable called “C” with levels 1, 2 as column 5
- Predictor 4: categorical variable called “D” with levels 1, 2, 3 as column 6

Refer to these data as “example.dat” and save it under the current directory. We can fit logistic regression in R with the following codes.

```
### Read the data
d = read.table("example.dat")
names(d) = c("Yes", "No", "A", "B", "C", "D") # set var names
for(i in 3:6){ d[,i] = as.factor(d[,i]) } # convert to factor
# Alternatively, this can be done by
d = read.table("example.dat", colClasses=c("integer","integer",
               "factor","factor","factor","factor"))
names(d) = c("Yes", "No", "A", "B", "C", "D")
### Fit logistic regression via glm
# main effect only
fit1 = glm(cbind(Yes, No) ~ A + B + C + D, data = d, family = binomial())
# main effect & up to 2-way interactions
fit2 = glm(cbind(Yes, No) ~ (A + B + C + D) * (A + B + C + D),
           data = d, family = binomial())
# Recall that A * B <=> A + B + A:B, main effect plus interaction
# A : B : C     means three-way interaction term only
# A * B * C     means main effect, all 2-way plus 3-way interaction
```

7.6 PROBLEM SET 2

1. For the Gamma distribution,

   (a) Write the parameters (μ, ν) as functions of parameters (λ, κ), using the two parametrizations given in section 4.3 of the notes.

   (b) Show the Gamma density belongs to the exponential family, write it in the canonical exponential form and identify the parameters.

   (c) Use the exponential form to obtain expectation and variance.

   (d) Derive the variance function.

   (e) Derive the canonical link.

2. For the inverse Gaussian distribution, carry out parts (b)-(e) of the previous problem.

3. (a) For binomial and Gamma models, obtain the deviance.
   (b) For these models, derive the Pearson and Deviance residuals.

4. (a) Repeat the Poisson regression fit for the Melanoma data and obtain Pearson and deviance residuals.
   (b) Compare these two types of residuals by explorative analysis (summary statistics, box plots and plotting them against the fitted value). Describe your findings.
   (c) Is there any indication of lack of fit when checking the residual plots?
   (d) Obtain the deviance table for the two predictors ("ANOVA table").
   (e) Based on the table, do a quick stepwise variable selection (not running the program again).
   (f) Select the predictors by AIC. Discuss your results and determine your final model. What do you conclude about the incidence of melanomas, based on your final model?

5. Extend (3.4) to the fitting of surfaces to data of the type $(X_{i1}, X_{i2}, Y_i)_{i=1,\dots,n}$, where X_{i1}, X_{i2} are first resp. second component of a 2-dimensional predictor vector of continuous predictors. Can you extend this to the case of $p-$dimensional predictors? Discuss the performance of the resulting estimates for the case where p is large, either theoretically or through a small simulation study.

6. Extend (3.4) to the estimation of the first derivative $g^{(1)}(x)$ by locally fitting a quadratic polynomial.

   (a) Describe your proposed estimator in a way that others can also implement it (do not provide code).
   (b) How would you choose the bandwidth? Provide at least two methods and discuss how you would choose among them.
   (c) Apply the estimator with bandwidth choice to the zmk data.

7. In the nonparametric regression model

$$Y_i = g(\frac{i}{n}) + e_i, \quad i = 1, \dots, n, \;\; e_i \quad \text{i.i.d.}, \quad E(e_i) = 0, \text{var}(e_i) = \sigma^2,$$

consider the linear estimators

$$\hat{g}(x) = \sum_{i=1}^{n} W_i(x) Y_i.$$

(a) Show that these estimators will be asymptotically unbiased if: (i) g is Lipschitz continuous on its domain, i.e., for any x_1, x_2 in the domain of g, one has that $|g(x_1) -$

$g(x_2)| \leq L|x_1 - x_2|$ for a constant $L > 0$; (ii) $W_i(x) = 0$ if $|\frac{i}{n} - x| \geq h(n)$ for a sequence $h(n) \to 0$ as $n \to \infty$; (iii) $\sum_{i=1}^n W_i(x) = 1$; and (iv) $W_i(x) \geq 0$ for all i. Hint: Use bounds for $|g(x) - g(\frac{i}{n})|$.

(b) Obtain $\text{var}(\hat{g}(x))$ for finite sample size n.

(c) For which choice of the weights $W_i(x)$ is the variance minimized if the bias condition $\sum_{i=1}^n W_i(x) = 1$ is required? For a quotient type a kernel estimator, which kernel would minimize the variance? Would you choose this kernel in practical applications?

8. Show that (4.4) indeed holds for cumulants.

9. Use the cumulant generating property of the function $b(\theta)$ to obtain the first three central moments of the binomial distribution.

8 Obtaining MLEs and information

8.1 Review of numerical ML estimation

We aim at maximizing $l(\beta)$ w.r. to $\beta \in \mathcal{R}^p$.

Score vector	$U(\beta) = \begin{pmatrix} \partial l/\partial\beta_1 \\ \vdots \\ \partial l/\partial\beta_p \end{pmatrix}$
Hessian matrix	$H(\beta) = (\partial^2 l/\partial\beta_j\partial\beta_k),\ \ 1 \leq j,k \leq p$
Observed information	$I_{\text{obs}}(\beta) = -H(\beta)$
Information	$I(\beta) = E(I_{\text{obs}}(\beta))$
ML estimating equations	$U(\hat{\beta}) = 0, \qquad$ root $\hat{\beta}$ is MLE

Newton-Raphson iteration:

Expand score function by a multivariate Taylor expansion,

$$0 = U(\hat{\beta}) = U(\beta) + H(\beta)(\hat{\beta} - \beta) + O(\| \hat{\beta} - \beta \|^2) \tag{8.1}$$

$$\hat{\beta} = \beta - H^{-1}(\beta)U(\beta) + O(\cdots) \leftarrow \text{negligible under regularity conditions}$$

This motivates the Newton-Raphson iteration

$$\begin{array}{rcl} \textit{Initialization } \hat\beta_{(0)} & = & \text{starting value} \\ & = & \text{(for GLM) regular least squares estimate for } \beta \text{ from multiple regression} \end{array}$$

Updating step

$$\hat\beta_{(l+1)} = \hat\beta_{(l)} - H^{-1}(\hat\beta_{(l)})U(\hat\beta_{(l)}) \tag{8.2}$$

For acceleration of convergence, often use acceleration factors for updating (Levenberg-Marquardt). An important variant is *Fisher Scoring*, where we replace $I_{obs} = -H$ by I, the expected information:

$$\hat\beta_{(l+1)} = \hat\beta_{(l)} + I^{-1}(\hat\beta_{(l)})U(\hat\beta_{(l)}). \tag{8.3}$$

According to the second Bartlett identity, approximately,

$$\operatorname{cov}(U(\beta)) = I(\beta).$$

Assume the limiting information $I_\infty(\beta) = \lim_{n\to\infty} n^{-1}I(\beta)$ is well-defined and is invertible. By the law of large numbers, then also $n^{-1}I_{obs}(\beta) \to I_\infty(\beta)$ in probability. Then, using the score statistic (2.6),

$$nI_{obs}^{-1}(\beta) \to_P I_\infty^{-1}(\beta), \quad n^{-1/2}I^{1/2}(\beta) \to I_\infty^{1/2}(\beta), \quad I^{-1/2}(\beta)U(\beta) \to_D N_p(0, I_p).$$

From (8.1), applying Slutsky's theorem and ignoring terms of smaller order,

$$\begin{array}{rcl} \sqrt{n}(\hat\beta - \beta) & = & \sqrt{n}[-H^{-1}(\beta)U(\beta)] \\ & = & \sqrt{n}I_{obs}(\beta)^{-1}U(\beta) \\ & = & [nI_{obs}(\beta)^{-1}]\ [n^{-1/2}I^{1/2}(\beta)]\ [I^{-1/2}(\beta)U(\beta)] \\ & & \to_D N_p(0, I_\infty^{-1}(\beta)). \end{array} \tag{8.4}$$

This result provides asymptotic normality for the MLEs and justifies the finite sample approximations

$$\hat{\beta} \sim_{approx.} N_p(\beta, I^{-1}(\hat{\beta})), \quad \hat{\beta} \sim_{approx.} N_p(\beta, I_{obs}^{-1}),$$

which are used for finite-sample inference, corresponding to the Wald statistic (2.7).

8.2 Application to GLMs

Define $n \times n$-diagonal *weight matrix*

$$W = \{\mathrm{diag}(g'(\mu_1)^2\, V(\mu_1)\, \phi, \dots, g'(\mu_n)^2\, V(\mu_n)\, \phi)\}^{-1} = (w_{ij})_{1 \le i,j \le n}.$$

where $\eta_i = g(\mu_i)$, $\eta = g(\mu)$, $\mu = g^{-1}(\eta)$, and $V(\mu_i)$ is the *variance function*, ϕ is a constant. Further define the $n \times p$ *design matrix*

$$X = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

In the GLM, the log likelihood for one observation is $l(y, \theta, \phi) = \{(y\theta - b(\theta))/\phi + c(y, \phi)\}$, and the r-th element of the score vector $U = (u_1, \dots, u_p)^\top$ (for one observation and in terms of β) is

$$u_r = \frac{\partial l}{\partial \beta_r} = \frac{\partial l}{\partial \theta} \frac{\partial \theta}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \beta_r} \tag{8.5}$$

by the chain rule applied to the composite mapping

$$\beta \to \eta \to \mu \to \theta \to l, \qquad \text{where}$$

$$\eta = X\beta, \ \ \mu = g^{-1}(\eta), \ \ \theta = b'^{-1}(\mu) \qquad \text{in the exponential family.}$$

Identifying the derivatives, denoting by x_r the r-th column of the design matrix,

$$\begin{aligned}
\frac{\partial \eta}{\partial \beta_r} &= x_r \\
U &= \frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{\phi} = \frac{y-\mu}{\phi}, \text{ since } \mu = b'(\theta) \\
\frac{\partial \theta}{\partial \mu} &= \frac{\partial b'^{-1}(\mu)}{\partial \mu} = \frac{1}{b''(b'^{-1}(\mu))} = \frac{1}{V(\mu)},
\end{aligned} \tag{8.6}$$

noting $V(\mu) = V(\theta(\mu)) = b''(\theta(\mu)) = b''(b'^{-1}(\mu))$ and therefore

$$u_r = \frac{y-\mu}{\phi V(\mu)} \frac{\partial \mu}{\partial \eta} x_r = (y-\mu) W \frac{d\eta}{d\mu} x_r. \tag{8.7}$$

8.3 Estimating equation

$$\sum_{i=1}^{n} u_{ri} = \sum_{i=1}^{n} (\frac{\partial l_i}{\partial \beta_r}) = 0$$

$$\Rightarrow \sum_{i=1}^{n} w_{ii}(y_i - \mu_i) \frac{d\eta_i}{d\mu_i} x_{ir} = 0$$

Hessian: $H = (h_{rs})_{1 \le r,s \le p}$ with elements

$$\begin{aligned}
h_{rs} &= \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \beta_r \partial \beta_s} = \sum_{i=1}^{n} \frac{\partial u_{ri}}{\partial \beta_s} \\
&= \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial}{\partial \beta_s} (w_{ii} \frac{d\eta_i}{d\mu_i}) x_{ir} + \sum_{i=1}^{n} w_{ii} \frac{d\eta_i}{d\mu_i} x_{ir} \frac{\partial}{\partial \beta_s} (y_i - \mu_i)
\end{aligned}$$

$$H = X^\top \left[\frac{\partial}{\partial \beta_s} W \frac{d\eta}{d\mu} \right] (y - \mu) \; - \; X^\top W X, \tag{8.8}$$

since

$$\frac{\partial}{\partial \beta_s} (y_i - \mu_i) = -\frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_s} = -\frac{\partial \mu_i}{\partial \eta_i} x_{is}. \tag{8.9}$$

From $E\,y_i = \mu_i$, we obtain

$$I = -E(H) = X^\top W\, X \tag{8.10}$$

For the case of a canonical link function, $\theta = \eta$, for one observation, using $\partial\theta/\partial\eta \equiv \text{id}$,

$$\theta = \eta \quad \Rightarrow_{(8.5)} \quad u_r = \frac{\partial l}{\partial \beta_r} = \frac{\partial l}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta_r} = \frac{\partial l}{\partial \theta}\frac{\partial \eta}{\partial \beta_r} = x_r \frac{y-\mu}{\phi}$$

$$h_{rs} = \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \beta_r \partial \beta_s} = \sum_{i=1}^{n} \frac{\partial u_{ri}}{\partial \beta_s} =_{(8.9)} \frac{1}{\phi}\sum_{i=1}^{n}(-\frac{\partial \mu_i}{\partial \eta_i} x_{is} x_{ir}) = (-X^\top W\, X)_{rs} \tag{8.11}$$

as

$$w_{ii} = \frac{1}{g'(\mu_i)^2 \phi V(\mu_i)} = \frac{1}{\phi}\frac{\partial \theta_i}{\partial \mu_i}(\frac{\partial \mu_i}{\partial \eta_i})^2 = \frac{1}{\phi}\frac{\partial \mu_i}{\partial \eta_i}$$

in the canonical case, using $\frac{\partial \theta_i}{\partial \mu_i} = 1/V(\mu_i)$, by (8.6). This implies

$$I = I_{\text{obs}} = X^\top W\, X. \tag{8.12}$$

Conclusion: For the canonical link ($\theta = \eta$), , $I_{\text{obs}} = I$ and Fisher scoring is the same as Newton-Raphson with observed information. For any link, the Fisher information in a GLM is $I = X^\top W\, X$.

9 Iterated weighted least squares

Define auxiliary variables

$$Z = \eta + (y-\mu)\frac{d\eta}{d\mu} = \begin{bmatrix} \eta_1 + (y_1 - \mu_1) g'(\mu_1) \\ \vdots \\ \eta_n + (y_n - \mu_n) g'(\mu_n) \end{bmatrix} \qquad n \times 1 \text{ - vector}$$

Iterated weighted least squares algorithm to obtain MLE $\hat{\beta}$:

1. Starting value for β, Z, W: $\hat{\beta}_{(0)}$, often obtained by regular (unweighted) multiple regressions: $Z_{(0)} = g(Y)$; $W_{(0)} = I_{n\times n}$, where $I_{n\times n}$ is the identity matrix.

2. Updating step: Given $\hat{\beta}_{(l)}$, obtain $\hat{\beta}_{(l+1)}$ by a weighted multiple linear regression step:

$$\begin{aligned} \hat{\beta}_{(l+1)} &= \text{argmin}_\beta \sum_{i=1}^{n} \{Z_{(l)i} - \sum_{j=1}^{p} x_{ij}\beta_{(l)_j}\}^2 W_{(l)ii} \\ &= \text{argmin}_\beta (Z_{(l)} - X\beta)^\top W_{(l)}(Z_{(l)} - X\beta), \qquad (9.1) \end{aligned}$$

   then update

$$\begin{aligned} \hat{\eta}_{(l+1)} &= X\hat{\beta}_{(l+1)}, \\ \hat{\mu}_{(l+1)} &= g^{-1}(\hat{\eta}_{(l+1)}), \\ W_{(l+1)} &= \{\text{diag}(g'(\hat{\mu}_{1(l+1)})^2 V(\hat{\mu}_{1(l+1)})\hat{\phi}, \ldots, g'(\hat{\mu}_{n(l+1)})^2 V(\hat{\mu}_{n(l+1)})\hat{\phi})\}^{-1}, \\ Z_{(l+1)} &= \hat{\eta}_{(l+1)} + (y - \hat{\mu}_{(l+1)})\frac{d\eta}{d\mu}|_{\mu=\hat{\mu}_{(l+1)}}. \end{aligned}$$

   Once β is updated, then update η, μ, W and Z.

3. Stopping criterion: Stop the iteration, if

$$\frac{\|\hat{\beta}_{(\ell+1)} - \hat{\beta}_{(\ell)}\|}{\|\hat{\beta}_{(\ell)}\|} \leq \varepsilon$$

   for a suitably chosen small constant ε, or if the number of iterations is larger than a pre-specified number N.

Motivation: Note that for

$$Z = \eta + (y - \mu)g'(\mu) \approx g(y),$$

one has

$$Z - \eta = (y - \mu)g'(\mu)$$

and

$$(Z-\eta)^2\frac{1}{g'(\mu)^2V(\mu)\phi} = \frac{(y-\mu)^2}{V(\mu)\phi} = P,$$

where the l.h.s. can be rewritten as $(Z-X\beta)^\top W(Z-X\beta)$. Therefore, weighted least squares aims at minimizing the Pearson distance P. Note that $\hat{\phi}$ may or may not be updated.

Analyzing this iteration, we obtain

$$\hat{\beta}_{(l+1)} = (X^\top\, W_{(l)}\, X)^{-1}X^\top\, W_{(l)}\, Z_{(l)} \tag{9.2}$$

as the weighted least squares solution from normal equations.
Inserting $Z_{(l)}$, noting that by (8.7)

$$U(\hat{\beta}) = X^\top W(y-\hat{\mu})\frac{\partial\eta}{\partial\mu},$$

$$\begin{aligned}
\hat{\beta}_{(l+1)} &= (X^\top\, W_{(l)}\, X)^{-1}X^\top\, W_{(l)}(X\hat{\beta}_{(l)} + (y-\hat{\mu}_{(l)})\,\frac{d\eta}{d\mu}|_{\mu=\hat{\mu}_{(l)}}) \\
&= \hat{\beta}_{(l)} + (X^\top W_{(l)}X)^{-1}X^\top W_{(l)}(y-\hat{\mu}_{(l)})\frac{d\eta}{d\mu}|_{\mu=\hat{\mu}_{(l)}} \\
&= \hat{\beta}_{(l)} + (X^\top\, W_{(l)}\, X)^{-1}U(\hat{\beta}_{(l)}) = \hat{\beta}_{(l)} + I^{-1}U(\hat{\beta}_{(l)}).
\end{aligned}$$

Result: For ML estimation in the GLM, Fisher scoring and iterated weighted least squares are equivalent. For the canonical link, Fisher scoring and Newton-Raphson iteration with observed information are equivalent.

Notes: (1) Inference is usually based on expected information

$$E(I_{obs}) = I = X^\top\, W\, X,$$

and on the approximate distribution

$$\hat{\beta} \sim N_p(\beta, (X^\top W X)^{-1}), \quad W = W(\beta) \tag{9.3}$$

for large samples, from which we obtain simultaneous inference or inference for individual components as described previously.

(2) The linearization through IWLS is a key device to extend arguments and tools that are available for (weighted) least squares to GLMs.

(3) An underlying regularity condition for the asymptotics to work is:

$$I_\infty = \lim_{n\to\infty} \frac{1}{n}(X^\top W X)$$

exists and is invertible. This being an asymptotic condition, it has no direct bearing on the finite sample situation.

Example for inference: To test

$$H_0: \ A\beta = \zeta, \text{ where } \quad A \text{ is } q \times p, \ \beta \in \mathcal{R}^p \text{ and } \zeta \in \mathcal{R}^q,$$

note that by (9.3), approximately,

$$A\hat{\beta} \sim N_q(A\beta, A(X^\top W X)^{-1} A^\top),$$

and that under H_0 we have, approximately,

$$(A\hat{\beta} - \zeta)^\top [A(X^\top W X)^{-1} A^\top]^{-1} (A\hat{\beta} - \zeta) \sim \chi^2_q,$$

if $A(X^\top W X)^{-1} A^\top$ is of full rank q. Can also construct confidence ellipsoid for parameter vector β, and for $\zeta = A\beta$, the latter based on the spectral decomposition of $[A(X^T W X)^{-1} A^T]$

(rather than that of $[X^T WX]^{-1}$).

Computational note: Implementing weighted least squares of Z on X with weight matrix W requires nothing more than a simple least squares step. Setting

$$X' = W^{1/2}X, \quad Z' = W^{1/2}Z, \tag{9.4}$$

the weighted least squares solution $\hat{\beta}$ is the same as the least squares solution for regressing Z' on predictors X'.

Note that the *hat matrix* H of the transformed regression is defined by $W^{1/2}\hat{Z} = HW^{1/2}Z$ for

$$H = W^{1/2}X(X^T WX)^{-1}X^T W^{1/2}. \tag{9.5}$$

Note that when applying (9.4), one has to do this for each iteration step as the matrices W are updated for each iteration. The hat matrix H to be used for diagnostics is the one obtained at the last iteration.

10 Models for binary data

10.1 Binomial regression

Main applications in Biostatistics:

Epidemiology:	Case-control studies, Cohort studies
Dose-Response:	Bioassay
Bioinformatics:	Discriminant analysis, "Supervised learning"
	Cluster analysis, "Unsupervised learning"

Data (x_i, y_i), $x_i \in \mathcal{R}^p$ covariate vector for response $y_i \in \{0, 1\}$ for i-th subject, $\;i = 1, \ldots, n$, y_i and y_j are independent for $i \neq j$.

Bernoulli experiment with response probability π_i:

$$P(y_i = 1) = \pi_i = \pi(x_i)$$

$$P(y_i = 0) = 1 - \pi_i = 1 - \pi(x_i)$$

Often, different subjects have the same covariate levels, in which case we have *Grouped Data*:

Covariate levels	No. of subjects at level	No. of positive responses (out of n_i)
x_1	n_1	y_1
$\vdots$	$\vdots$	$\vdots$
x_m	n_m	y_m

Total number of subjects $n = \sum\limits_{i=1}^{m} n_i$. The case $m = n$ with $n_i = 1$ corresponds to ungrouped data.

10.2 Link functions

In this case, $\mu = \pi \in [0, 1]$, so need link function

$g : \ \mu \to \ \eta = X\beta \qquad \mu \in [0, 1], \quad \eta \in \mathcal{R}$

$g^{-1} : \ \mathcal{R} \to [0, 1]$, strictly monotone. Any cdf $F : \ \mathcal{R} \to [0, 1]$ for which a pdf $f = F'$ exists with $f(x) > 0$ for all x is a possible choice:

$$g^{-1} = F, \quad g = F^{-1}.$$

Common choices are:

logit: $g^{-1}(x) = \text{expit}(x) = \frac{\exp(x)}{1+\exp(x)}, \quad x \in \mathcal{R},$ logistic cdf.

$g(\pi) = \log(\frac{\pi}{1-\pi}) = \text{logit}(\pi), \quad \pi \in [0, 1].$

This is the canonical link $\Rightarrow$ *Logistic regression model*.

probit: $g^{-1}(x) = \Phi(x), \ x \in \mathcal{R}, \ \Phi$ Gaussian cdf.

$g(\pi) = \Phi^{-1}(\pi), \ \pi \in [0, 1]$ which often yields results that are close to logit.

complementary log-log: $g(\pi) = \log(-\log(1 - \pi))$, asymmetric link function.

10.3 Likelihood

For grouped data, taking here y_i to be counts rather than fractions as before,

$$L = \prod_{i=1}^{m} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$$

$$l = \sum_{i=1}^{m} \Big\{ \frac{\frac{y_i}{n_i}[\log \pi_i - \log(1-\pi_i)] + \log(1-\pi_i)}{1/n_i} + \log \binom{n_i}{y_i} \Big\}$$

$$\theta = \text{logit}(\pi), \;\; b(\theta) = -\log(1-\pi) = \log(1+e^{\theta}), \;\; \phi_i = 1/n_i, \qquad (10.1)$$

using (5.1).

Simplified log likelihood, omitting terms not involving canonical parameters:

$$l(\pi, y) = \sum_{i=1}^{m} \{ y_i \log \frac{\pi_i}{1-\pi_i} + n_i \log(1-\pi_i) \}$$

10.4 Logistic regression

Logit link:

$$\eta = g(\mu) = \log \frac{\pi}{1-\pi} = \text{logit}(\pi), \quad \pi = \text{expit}(\eta) = \frac{e^{\eta}}{1+e^{\eta}}.$$

This is the canonical link, since $\theta = \text{logit}(\pi) = \eta$.

By (10.1),

$$\log(1-\pi_i) = -b(\theta_i) = -\log(1+\exp(\eta_i)),$$

we find

$$l(\pi, y) = \sum_{i=1}^{m} \{ y_i \sum_j x_{ij}\beta_j - n_i \log(1 + \exp(\sum_j x_{ij}\beta_j)) \}$$

$$
\begin{aligned}
u_r = \frac{\partial l}{\partial \beta_r} &= \underbrace{\sum_{i=1}^m y_i x_{ir}}_{\text{sufficient statistic}} - \sum_{i=1}^m n_i x_{ir} \underbrace{\frac{\exp(\sum_j x_{ij}\beta_j)}{1+\exp(\sum_j x_{ij}\beta_j)}}_{\text{expit}(\eta_i) = \pi_i} \\
&= \sum_{i=1}^m (y_i - n_i\pi_i)x_{ir}, \qquad \text{and}
\end{aligned}
$$

$$
h_{rs} = -\sum_{i=1}^m n_i x_{ir} \frac{\partial \mu_i}{\partial \beta_s} = -\sum_{i=1}^m n_i x_{ir} x_{is}(1-\pi_i)\pi_i, \tag{10.2}
$$

since with (5.3)

$$
\frac{\partial \mu_i}{\partial \beta_s} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_s} = \frac{\partial}{\partial \eta_i}\frac{e^\eta}{1+e^\eta} x_{is} = \pi_i(1-\pi_i)x_{is}
$$

We conclude that, as expected for the canonical link by the above general result, $\frac{\partial u_r}{\partial \beta_s} = h_{rs}$ does not depend on data, therefore $H = E\,H$.

10.5 Asymptotics

Two cases are of interest:

1. $n_i \to \infty$, m fixed, $\frac{n_i}{n} \to \lambda_i$, $\lambda_i > 0$, $\sum \lambda_i = 1$: "Grouped asymptotics".
2. $m \to \infty$, $n_i = n_1$ fixed, $n = mn_1 \to \infty$: "Level asymptotics", often used for the case $n_1 = 1$.

Regularity condition: $\frac{1}{n}(X^\top W X) \to I_\infty$, as $n \to \infty$, where I_∞ is positive definite, symmetric and invertible limit matrix. Then

$$
\sqrt{n}(\hat\beta - \beta) \to_D N_p(0, I_\infty^{-1}).
$$

10.6 Overdispersion in binary regression

Binomial variance function: $V(\pi) = \pi(1-\pi)$

For grouped data: $\text{var}(y_i) = n_i\pi_i(1-\pi_i),\ \ i = 1, \ldots, m$ (m levels).

Overdispersion: In practice, we find often that *actual dispersion* > *nominal dispersion*, i.e.,

$$\text{var}(y_i) > n_i\pi_i(1-\pi_i)$$

Then

$$\text{var}(y_i) = \sigma^2 n_i\pi_i(1-\pi_i),$$

where σ^2 is the *overdispersion parameter.*

Common causes for overdispersion:

- Parameter for each subject is not fixed but varies randomly around target value (*random effects model*).
- Data are clustered (neighborhoods, litters, families, cities).
- Data are correlated: Negative correlation within the responses for each group can lead to *underdispersion*, $\sigma^2 < 1$.

10.7 An analysis of clustered responses

Assume grouped data $(n_i, y_i),\ \ i = 1, \ldots, m$, where each group forms a "cluster" with its own random effect (say, they share the same "environment"): Response probability for i-th group p_i is random,

$$E(p_i) = \pi_i,\ \ \text{var}(p_i) = \tau\pi_i(1-\pi_i),\ \ \text{for a parameter}\, \tau,$$

$$Y_i|p_i \sim Bin(n_i, p_i).$$

Then

$$E(y_i) = E(E(y_i|p_i)) = n_i E(p_i) = n_i \pi_i,$$

$$\begin{aligned} \mathrm{var}(y_i) &= E[\mathrm{var}(y_i|p_i)] + \mathrm{var}[E(y_i|p_i)] \\ &= E(n_i p_i(1-p_i)) + \mathrm{var}(n_i p_i) \\ &= n_i\pi_i - n_i E p_i^2 + n_i^2 \tau \pi_i(1-\pi_i) \\ &= n_i\pi_i - [n_i \tau \pi_i(1-\pi_i) + n_i \pi_i^2] + n_i^2 \tau \pi_i(1-\pi_i) \\ &= n_i\pi_i(1-\pi_i)(1+(n_i-1)\tau) \end{aligned}$$

Now $\sigma_i^2 = 1 + (n_i - 1)\tau$ is the overdispersion parameter for the i-th level. Overdispersion occurs for the case $n_i > 1,\ \tau > 0$.

Estimation of the overdispersion parameter or scale factor via deviance:

$$\hat{\sigma}^2 = \frac{D}{n-p}.$$

10.8 Residuals and diagnostics

Assume m covariate levels, $\hat{y}_i = n_i\hat{\pi}_i = n_i\hat{\mu}_i,\ i = 1, \ldots, m,$

$$r_i = y_i - \hat{y}_i \qquad \text{residuals}$$

$$r_{P_i} = \frac{y_i - n_i\hat{\pi}_i}{(n_i\hat{\pi}_i(1-\hat{\pi}_i))^{1/2}}; \qquad P = \sum_{i=1}^{n} r_{P_i}^2 \quad \text{Pearson's } \chi^2 \text{ and Pearson residuals.}$$

With *weight matrix* W, from the weighted iterated least squares perspective,

$$W = \mathrm{diag}\{(g'(\mu_i)^2 V(\mu_i)\phi\}^{-1}$$

$$H = H(\beta) = W^{1/2}X(X^\top W X)^{-1}X^\top W^{1/2} = (h_{ij})_{1\le i,j\le n}$$

.

This is the *hat matrix* as in (9.5): $H^2 = H$ and $H^T = H$, ie., H is a (orthogonal) projection matrix with $\hat{Z} = HZ$ and $\text{Trace}(H) = \sum_{i=1}^n h_{ii} = p$.
For diagnostics, one uses the matrix H at the last IWLS iteration.

Leverage points are suspected if $h_{ii} > 2p/n$.
To detect outliers/influential observations:
Obtain *Cook's distance* for the i-th subject (experiment), a measure of the influence of the i-th subject on the regression relation:

$$D_i = \frac{(\hat{y}_i - \hat{y}^{(-i)})^T(\hat{y}_i - \hat{y}^{(-i)})}{p\tilde{\sigma}^2} = \frac{e_i^2}{p\tilde{\sigma}^2}\frac{h_{ii}}{(1-h_{ii})^2}.$$

Here, h_{ii} is the i-th diagonal element of H, the hat matrix that is obtained at the last iteration, and $\tilde{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^n e_i^2$, where $e_i = Z_i - \hat{Z}_i$ is the (standard) residual when regressing Z on the predictors (see section 9).

For deviance residuals, Pearson residuals etc., see above. Plot residuals vs individual components x_i, or vs linear predictors $\hat{\eta}_i$.

Checking the link function: Residuals should not show trend in the mean. Especially in case of ungrouped data, they will show apparent patterns which are not cause for concern. Various smoothers can be used to check whether smoothed residuals are close to zero throughout.

Smoothing helps to check goodness-of-link and goodness-of-fit of the variance function. Especially for the case of ungrouped data, where $m = n$, $n_i = 1$, grouping the data artificially by binning or applying a smoothing method to the residual plots is often necessary.

10.9 Link function choice by an embedding technique

Assume link function belongs to a parametrized family,

$$g(\pi_i, \alpha) = \eta_i, \;\; \eta_i \text{ linear predictor.}$$

The correct link is assumed to be in this family, e.g., it is attained for $\alpha = \alpha_0$.
Aranda-Ordaz link family:

$$g(\mu_i, \alpha) = g(\pi_i, \alpha) = \log\{\frac{(1-\pi_i)^{-\alpha} - 1}{\alpha}\} \tag{10.3}$$

$$\begin{aligned} \alpha = 1: \quad & g(\pi_i, 1) = \log \frac{\pi_i}{1-\pi_i} = \text{logit}(\pi_i) \\ \alpha \to 0: \quad & \frac{(1-\pi_i)^{-\alpha} - 1}{\alpha} \to \log \frac{1}{1-\pi_i} \quad \text{as} \\ & \lim_{\alpha \to 0} \frac{x^\alpha - 1}{\alpha} = \log x \end{aligned}$$

by L'Hôpital's rule, noting $\frac{d}{d\alpha} \exp(\alpha \log(x)) = x^\alpha \log(x)$. Then

$$g(\pi_i, 0) = \log(-\log(1-\pi_i))$$

is the complementary log link.

Goodness-of-fit check for assumed model $\alpha = \alpha_0$ (usually $\alpha_0 = 1$, logistic model, in the binomial regression situation): Observe

$$g(\pi_i, \alpha) \approx g(\pi_i, \alpha_0) + (\alpha - \alpha_0) \frac{\partial g(\pi_i, \alpha)}{\partial \alpha}|_{\alpha=\alpha_0} = \eta_i + (\alpha - \alpha_0)\gamma_i \leftarrow \text{covariate}$$

Testing whether additional term $(\alpha - \alpha_0)\gamma_i$ is relevant then corresponds to testing the null hypothesis $H_0: \;\; \alpha = \alpha_0 \; (\alpha_0 \neq 0)$.

For g as in (10.3), and first obtaining fits $\hat{\pi}_i$ by choosing $\alpha = \alpha_0$,

$$\gamma_i = \frac{\partial g}{\partial \alpha}|_{\alpha=\alpha_0} = \frac{\log(1-\hat{\pi}_i)}{(1-\hat{\pi}_i)^{\alpha_0}-1} - \frac{1}{\alpha_0}; \tag{10.4}$$

for the special case $\alpha_0 = 1$,

$$\gamma_i = -[1 + \hat{\pi}_i^{-1}\log(1-\hat{\pi}_i)]. \tag{10.5}$$

This will be added as an additional covariate, then the test for $H_0 : \alpha = \alpha_0$, i.e., whether the parameter θ for γ_i is significant, serves as ”goodness-of-link test”.

If the goodness-of-link test rejects, we need to look for another link function. For this, one can use the fitted parameter $\hat{\theta} = \hat{\alpha} - \alpha_0$, then update $\hat{\alpha} = \alpha_0 + \hat{\theta}$ and use the link function $g(\mu, \hat{\alpha})$.

10.10 PROBLEM SET 3

1. For the Poisson regression model with log link and p predictor variables (including the intercept), derive the score vector, as well as observed and expected information matrix, in a similar way as the derivation for logistic regression that is given in Section 10.4.

2. Provide details of the derivation leading to (8.10) and (8.12), i.e., follow the notes and write out the details step by step.

3. Prove (9.2).

4. Prove that the weighted least squares updating step in the iterated weighted least squares algorithm can be implemented by unweighted least squares for the linearly transformed variables $X' = W^{1/2}X,\ Z' = W^{1/2}Z$.

5. Show that the hat matrix H in (9.5) indeed is a projection matrix for an orthogonal projection, as is expected for a hat matrix.

6. Derive the weight matrix and the information matrix for the following models: (a) Logistic regression with the canonical link (b) Poisson regression with the canonical link.

Note: You need to organize your output and write a brief report for all of the following as well as subsequent data analysis problems.

7. The data set chemo.dat contains the number of epileptic seizures during four two-week periods for n=59 patients in col 2 - col 5. Col. 6 indicates whether an active treatment (0=placebo, 1=active) was given. Col. 7 is a baseline rate for seizures, established before the trial started, Col 8 contains the age of a patient. Investigate a Poisson model, defining col 2 + col 3 + col 4 + col 5 as the dependent variable.

   Before you start the analysis, use basic scatterplots to identify one subject as outlier, then remove this outlier and work with the remaining 58 subjects. Predictors are baseline, age and treatment. Use the canonical link. Include all predictors in your model. Check and report goodness-of-fit using deviance and Pearson residuals and interpret your findings. Which is the inference one would be interested in? Please carry out the resulting test and interpret all your findings.

8. For the Pima Indians diabetic study, fit a binary regression model with canonical link. (a) Obtain the fit with all predictors in the model, carry out the test for overall regression effect and perform diagnostics. (b) Select the relevant predictors and thus the final model by AIC and BIC. Compare these model selectors. (c) Choose one final model, obtain inference and interpret your findings.

9. Fit a logistic regression model to the data set lung.dat. Byssinosis is a lung disease of cotton workers and is related to the dustiness of the workplace. The data set includes the response byssinosis yes (number of cases), byssinosis no (number of cases), dust level (coded 1-3), race, sex, smoking, length of employment (coded 1-3). Each row of the data corresponds to one workplace. The coding of the covariates is as follows:

   Dust: Dustiness of the Workplace (1 high, 2 medium, 3 low)

   Race: Ethnic Group of Worker (1 white, 2 other)

   Smoking: Smoking Status (1 smoker, 2 non-smoker)

   Sex: (1 male, 2 female)

   Length of employment: (1 less than 10 years, 2 10-20 years, 3 over 20 years.)

   (a) Use the deviance table and AIC to select relevant predictors. Compare these models. For the smaller of the models, plot deviance and Pearson residuals and compare them. Interpret your final model. For the predictors coded 1-3, do not introduce indicator variables but rather use the given levels as numerical values with an unknown unit.

   (b) For which predictors would you consider one-sided tests? Provide your reasons and write the null hypotheses and alternatives and then use your output from

(a) to carry out these tests. Note: Your considerations about one-sided testing should be a priori and not influenced by the outcomes of your model fits in (a).

(c) For your selected model, obtain Cook's distance and leverage plots and interpret the results.

(d) Investigate whether there are interactions between the predictors. Write the overall null hypothesis for no interactions in the form $H_0 : A\beta = 0$ and determine the df. Select the interactions which you believe should be included in the final model. Provide your reasons and interpret your model, taking into account both main effects and interactions.

11 Multinomial regression models

11.1 Exponential family representation

Relating *categorical outcomes* to covariates: If one has just two categories (e.g., patient has schizophrenia, yes or no) then use binomial regression. However, often outcomes are classified into more than two outcome *categories*. Epidemiologists often form categories of initially continuous data, e.g., Age → Young, Middle-aged, Old; Blood pressure → normal, elevated, strongly elevated. The cut-points that define the boundaries between categories may come from background knowledge; or one can use quantiles, e.g., Q1, Q2, Q3 if converting continuous data into 4 categories.

Nominal Outcomes: Not ordered, such as disease type. Examples: Gender (m,f,other), Blood type, Color.

Ordinal Outcomes: E.g., no, low level, high level of depression or other psychiatric symptoms (3 levels); no, little, medium, lots, unbearable pain (5 levels); Normal, Mild Cognitive Impairment, Alzheimer's (3 levels of mental impairment); many questionnaires have 5 levels: do not agree, mostly disagree, neither agree nor disagree, mostly agree, agree.

One option to handle such data (and also continuous data): Dichotomization, combining categories until only two are left, followed by analysis with a (much simpler) binary regression model (often *median dichotomization* when starting with continuous variables). Then implement a binary regression model.

For categorical outcomes with M levels, define indicators Y_{im} with $Y_{im} = 1$ if for the i-th subject an outcome in category m is observed, $m = 1, \ldots, M-1$.
One needs one less indicator than categories, as usual consider carefully where to place the baseline category, defined by $Y_{im} = 0$, $m = 1, \ldots, M-1$ and corresponding to the M-th category.

Responses are $(M-1)$-vectors Y_i with $\sum_{m=1}^{M-1} Y_{im} = 1 \quad \text{or} \quad = 0:$

$$Y_i \sim Mult(\pi_i, 1), \quad \pi_i = (\pi_{i1}, \ldots, \pi_{iM-1})^\top, \quad 0 \le \pi_{im} \le 1, \quad \sum_{m=1}^{M-1} \pi_{im} \le 1,$$

the *Multinomial Distribution* with parameters $\pi_i = (\pi_{i1}, \ldots, \pi_{iM-1})^\top$, where the response probabilities $\pi_i = \mu_i$ form a $(M-1)$-vector of means that are related to linear predictors η_i.
Defining

$$\pi_{iM} = 1 - \sum_{m=1}^{M-1} \pi_{im},$$

the pmf is

$$\begin{aligned} P(Y_i = y_i) &= \pi_{i1}^{y_{i1}} \pi_{i2}^{y_{i2}} \ldots \pi_{i,M-1}^{y_{i,M-1}} \pi_{iM}^{(1-\sum_{m=1}^{M-1} y_{im})} \\ & E(Y_{im}) = \pi_{im}, \quad \text{cov}(Y_{im}, Y_{ij}) = c_{ijm}, \end{aligned} \qquad (11.1)$$

where for $1 \le m \le M-1$,

$$c_{ijm} = \begin{cases} -\pi_{ij}\pi_{im} & \text{if} \quad j \ne m \\ \pi_{im}(1-\pi_{im}) & \text{if} \quad j = m. \end{cases} \qquad (11.2)$$

For the case of n_i repeated measurements obtained at the i-th covariate level, the model can be written in an equivalent aggregated form. Define Z_{im} to be the aggregated counts of observations for category m (not using indicators) and consider vectors of counts $Z_i = (Z_{i1}, \ldots, Z_{iM})^\top$. The pmf can then be written as

$$P(Z_i = z_i) = \frac{n_i!}{z_{i1}! \ldots z_{iM}!} \pi_{i1}^{z_{i1}} \ldots \pi_{iM}^{z_{iM}}, \tag{11.3}$$

in analogy to the binomial case. Here, $n_i = \sum_{j=1}^{M} z_{ij}$ is the number of subjects that share the same covariate levels, analogous to the grouped binomial case.

Using indicator vectors Y_i, obtain exponential family representation with a vector canonical parameter θ_i of dimension $(M-1)$:

$$\begin{aligned}
P(Y_i = y_i) &= \pi_{iM}^{(1-\sum_{m=1}^{M-1} y_{im})} \prod_{m=1}^{M-1} \pi_{im}^{y_{im}} \\
&= \exp\left(\sum_{m=1}^{M-1} y_{im} \log(\pi_{im}) + (1 - \sum_{m=1}^{M-1} y_{im}) \log(\pi_{iM}) \right) \\
&= \exp\left(\sum_{m=1}^{M-1} y_{im} \log(\pi_{im}/\pi_{iM}) + \log(\pi_{iM}) \right) \\
&= \exp\left(y_i^\top \theta_i - b(\theta_i) \right)
\end{aligned}$$

with

$$\theta_{im} = \log(\pi_{im}/\pi_{iM}), \quad b(\theta_i) = -\log(\pi_{iM}) = \log[1 + \sum_{m=1}^{M-1} \exp(\theta_{im})].$$

Note that

$$\frac{\partial}{\partial \theta_{im}} b(\theta_i) = \pi_{im}, \tag{11.4}$$

$$\frac{\partial^2}{\partial \theta_{im} \partial \theta_{ij}} b(\theta_i) = \begin{cases} -\pi_{ij}\pi_{im} & \text{if} \quad j \neq m \\ \pi_{im}(1 - \pi_{im}) & \text{if} \quad j = m, \end{cases} \tag{11.5}$$

as would be expected from the properties of the cumulant generating function b.

11.2 Multinomial regression

Proportional Odds Model

In a multinomial regression model with response categories coded as $j = 1, 2, \ldots, M$, define new responses z_{im} with $z_{im} = \sum_{j=1}^{m} y_{ij}$.

Then $z_{im} = 1$ if $y_{ij} = 1$ for a $j \le m,\ 1 \le m \le M - 1$, otherwise $z_{im} = 0$.

With $\mu_{im} = E(z_{im})$ and binary link function g (most commonly chosen as logit link), the model is

$$g(\mu_{im}) = \beta_{0m} + X_i \beta,$$

with ordered intercepts $\beta_{01} \le \beta_{02} \le \ldots \le \beta_{0,M-1}, \quad \beta \in \Re^{p-1}$. This corresponds to a linear constraint and maximization of the likelihood is implemented subject to this constraint. The total number of parameters is $(p-1)+(M-1)$. This model is called proportional odds model for the case of a logit link.

Note: While this models works for both ordered and unordered categorical data, it is particularly useful if the data are ordered.

Baseline Odds Model

Here one first needs to select a baseline category. Assuming the baseline category is category 1, the linear predictor in this model is

$$\frac{\pi_{ij}}{\pi_{i1}} = \exp(\eta_{ij}), \quad \eta_{ij} = X_i \beta_j,\ \beta_j \in \Re^p, \quad 2 \le j \le M. \tag{11.6}$$

Equivalently, since for each i only one of the Y_{im} can satisfy $Y_{im} = 1$, and by using elementary conditional probability, (11.6) is equivalent to

$$\text{logit}\{P(Y_{ij} = 1 \mid Y_{ij} = 1 \text{ or } Y_{i1} = 1)\} = \eta_{ij}, \quad 2 \le j \le M, \tag{11.7}$$

and thus model (11.6) can be easily implemented as a series of logistic regressions. Since $1 = \sum_{j=1}^{M} \pi_{ij}$, (11.6) implies the more direct representations

$$\pi_{i1} = \frac{1}{1 + \sum_{j=2}^{M} \exp(\eta_{ij})}, \quad \pi_{ij} = \frac{\exp(\eta_{ij})}{1 + \sum_{j=2}^{M} \exp(\eta_{ij})} \quad \text{for} \quad 2 \leq j \leq M, \tag{11.8}$$

i.e., the estimates of the response probabilities can be easily recovered from the fitted linear predictors.

Disadvantages: This model has lots of parameters, their number is $(M-1)p$, while the proportional odds model only has $M+p-2$ parameters. Also a baseline needs to be selected and its choice may affect the interpretation of the results. All of these models can be fitted with standard software.

11.3 LAB 3: Multinomial Regression

11.3.1 Proportional Odds Model

The Proportional Odds version of multinomial regression performs logistic regression for the first i categories combined versus the rest, where $i = 1, \ldots, M-1$, with M denoting the total number of levels in the response. The model restricts the slope parameters to be the same for all $M-1$ submodels, and only the intercept coefficient increases in i. For *ordinal responses*, the Proportional Odds model provides a natural interpretation of the cumulative effect as response levels increase or decrease.

A good option to implement the Proportional Odds model is the **polr()** function in the MASS package:

```
library(MASS)
?polr
```

- Similar arguments required as in glm(): model formula, input data, etc.
- The response variable must be a *factor variable.*

- The option “method” specifies the desired link function to be used. Default is “logistic”, other choices include “probit”, “loglog”, “cloglog” and “cauchit”, “cauchit” is used for binary outcomes transforming to a heavy-tailed Cauchy distribution.
- Output: Summary table of coefficient estimates for all slopes and intercepts. Z-tests can be used to get approximate p-values for individual predictor effects.
- **Residuals have to be calculated based on observed response and fitted probabilities**, i.e. no built-in function available. (Consider how the residual plots look like for multinomial regression).

Example

```
library(MASS)
data(housing)
house.plr <- polr(Sat ~ Infl + Type + Cont, weights = Freq, data = housing)
summary(house.plr)
# n by M matrix of predicted prob:
prd_prob_po = predict(house.plr,housing,type=‘prob’)
prd_labl_po = predict(house.plr, housing) # vector of predicted labels
```

```
Call:
polr(formula = Sat ~ Infl + Type + Cont, data = housing, weights = Freq)

Coefficients:
                Value Std. Error t value
InflMedium     0.5664    0.10465   5.412
InflHigh       1.2888    0.12716  10.136
TypeApartment -0.5724    0.11924  -4.800
TypeAtrium    -0.3662    0.15517  -2.360
TypeTerrace   -1.0910    0.15149  -7.202
```

```
ContHigh       0.3603    0.09554   3.771
Intercepts:
            Value   Std. Error t value
Low|Medium  -0.4961  0.1248    -3.9739
Medium|High  0.6907  0.1255     5.5049

> head(prd_prob_po)
        Low    Medium      High
1 0.3784493 0.2876752 0.3338755
2 0.3784493 0.2876752 0.3338755
3 0.3784493 0.2876752 0.3338755
4 0.2568264 0.2742122 0.4689613
5 0.2568264 0.2742122 0.4689613
6 0.2568264 0.2742122 0.4689613
> head(prd_labl_po)
[1] Low  Low  Low  High High High
Levels: Low Medium High
```

11.3.2 Baseline Odds Model

The Baseline Odds version of multinomial regression is a combination of $M-1$ logistic regression models, targeting the conditional probabilities

$$\mathbb{P}(Y_{ij} = 1 | Y_{ij} = 1 \text{ or } Y_{i1} = 1),$$

where $j = 2, \ldots, M$, and level 1 denotes the baseline category.

For implementation, there are several functions in different packages in R to fit this model, such as **mlogit()** in the package **mlogit**. Here we use:

```
library(nnet)
?multinom
```

This does not require the data to be reshaped into a particular data type as mlogit does (input data frame needs to be an 'mlogit.data' object or equivalent).

Different from the proportional odds model, the baseline odds model regards one category as baseline level, and constructs submodels aiming at the conditional probabilities, where no restrictions on slopes or intercept terms are enforced. The total number of coefficients is $(M-1)\times p$, where p is the number of coefficients in each sub-model. It allows for much more freedom in the relationship between the response and predictors, but suffers from higher model complexity. Also, the choice of the baseline category can be arbitrary.

- Similar arguments as glm(): model formula, input data etc.
- The response variable can be a factor variable or a matrix of numbers of subjects falling in certain categories (aggregated format according to shared predictor levels).
- The summary table only provides coefficient estimates and corresponding standard errors. You can utilize the Gaussian approximation to perform simple hypothesis testing and obtain approximate p-values.
- Again, **residuals have to be calculated based on observed response and fitted probabilities**.

Example

```
library(nnet)
# here we use the previous example
housing.bo <- multinom(Sat ~ Infl + Type + Cont, data=housing, weights = Freq)
summary(housing.bo, digit=3)
prd_prob_bo = predict(housing.bo, type = 'prob')
head(prd_labl_bo)
prd_labl_bo = predict(housing.bo)
head(prd_labl_bo)
```

```
Call:
multinom(formula = Sat ~ Infl + Type + Cont, data = housing,
    weights = Freq)

Coefficients:
       (Intercept) InflMedium InflHigh TypeApartment TypeAtrium TypeTerrace ContHigh
Medium      -0.419      0.446    0.665        -0.436      0.131      -0.667    0.361
High        -0.139      0.735    1.613        -0.736     -0.408      -1.412    0.482

Std. Errors:
       (Intercept) InflMedium InflHigh TypeApartment TypeAtrium TypeTerrace ContHigh
Medium       0.173      0.142    0.186         0.173      0.223       0.206    0.132
High         0.159      0.137    0.167         0.155      0.211       0.200    0.124

> head(prd_prob_bo)
        Low    Medium      High
1 0.3955694 0.2601074 0.3443232
2 0.3955694 0.2601074 0.3443232
3 0.3955694 0.2601074 0.3443232
4 0.2602405 0.2674079 0.4723516
5 0.2602405 0.2674079 0.4723516
6 0.2602405 0.2674079 0.4723516
> head(prd_labl_bo)
[1] Low  Low  Low  High High High
Levels: Low Medium High
```

12 Models for count data

12.1 Poisson regression

Observe counts which depend on covariates:

No. of accidents, No. of earthquakes of a certain size per year, Counts of irregular heartbeats, Counts of T4-cells in HIV

"Classical" Poisson variance:

$$\mathrm{var}(y) = V(\mu) = \mu;$$

In the presence of overdispersion:

$$\mathrm{var}(y) = \sigma^2\mu,$$

where σ^2 is the overdispersion parameter.

If overdispersion is present, the Poisson regression model does not strictly apply, since the Poisson distribution does not include an extra parameter for overdispersion. In this case: Use quasi-Poisson or negative binomial model (see below).

Distribution: Pmf of model without overdispersion

$$P(Y = y) = e^{-\lambda}\lambda^y/y!, \ \ y = 0, 1, 2, \ldots$$

leads to the log-likelihood, omitting constants,

$$l(\mu, y) = \sum_{i=1}^{n}\{y_i \log \lambda_i - \lambda_i\}$$

and deviance

$$D(y,\lambda) = 2\{l(y,y) - l(\mu,y)\} = 2\sum_{i=1}^{n}(y_i \log(\frac{y_i}{\mu_i}) - (y_i - \mu_i))$$

12.2 Overdispersion in Poisson models

We would expect $\mathrm{var}(y) = E(y)$, but often find $\mathrm{var}(y) > E(y)$ (overdispersion) in Poisson models.

Classical mechanisms leading to overdispersion:

1. *The clustered Poisson process*

   $y = z_1 + \cdots + z_N, \;\; z_i \;\; iid, \;\; N \sim \text{Poisson}, \;\; N$ independent of $z_i's$.

   Examples:

   - Line transect sampling: $Z_i =$ no. of animals, animals sighted in clusters around transect line.
   - Insured accidents: N= no of accidents; $z_i =$ damage at i-th accident.

   Then, by independence of N and z_i,

$$\begin{aligned}
E(y) &= E(E(y|N)) = E(N(Ez_1)) = (EN)(Ez_1) \\
\mathrm{var}(y) &= E(\mathrm{var}(y|N)) + \mathrm{var}(E(y|N)) \\
&= E(N\mathrm{var}(z_1)) + \mathrm{var}(NEz_1) \\
&= (EN)\mathrm{var}(z_1) + (Ez_1)^2\mathrm{var}(N) \\
&= (EN)(Ez_1^2) \quad (\text{as } EN = \mathrm{var}(N), \;\; Ez_1^2 = \mathrm{var}(z_1) + (E(z_1))^2) \\
&> Ey \qquad \text{if } Ez_1^2 > Ez_1
\end{aligned}$$

2. *Mixture model*

   Example: No. of extrasystolic heartbeats Y recorded overnight. Diagnosis of arrhythmias in heartbeat requires long-term recording and automatic detection of extrasystolic heartbeats. Model:

$$Y \sim \text{Poisson}(Z)$$

The Poisson parameter Z will differ from subject to subject; extrasystolic beats are also recorded for healthy subjects.

Here, Z is a random subject effect, characteristic for an individual: Assume $Z \sim$ Gamma(κ, ρ) with pdf $f_{\kappa,\rho}(t) = \rho(\rho t)^{\kappa-1}e^{-\rho t}/\Gamma(\kappa)$,

$$E\,Z = \frac{\kappa}{\rho} = \mu, \;\; \mathrm{var}(Z) = \frac{\kappa}{\rho^2} = \frac{\mu}{\rho} \tag{12.1}$$

Then, setting $v = z(\rho+1)$, and plugging in pdf and pmf,

$$\begin{aligned} P(Y = y; \kappa, \rho) &= \int P(Y = y|Z = z) f_{\kappa,\rho}(z) dz \\ &= \int \frac{z^y e^{-z}}{y!} \rho(\rho z)^{\kappa-1} e^{-\rho z}/\Gamma(\kappa) dz \\ &= \frac{\rho^\kappa}{y!\Gamma(\kappa)} \int z^{y+\kappa-1} e^{-z(\rho+1)} dz \\ &= \frac{\rho^\kappa}{y!\Gamma(\kappa)(1+\rho)^{y+\kappa}} \int v^{y+\kappa-1} e^{-v} dv \\ &= \frac{\rho^\kappa \Gamma(y+\kappa)}{y!\Gamma(\kappa)(1+\rho)^{y+\kappa}} \\ &= \binom{y+\kappa-1}{\kappa-1} (\frac{\rho}{1+\rho})^\kappa (\frac{1}{1+\rho})^y, \end{aligned} \tag{12.2}$$

by definition of the Gamma function, $\Gamma(z) = \int v^{z-1}e^{-v}dv$, where generalized factorials are defined as $\Gamma(y) = (y-1)!$, an extension of $\Gamma(n+1) = n!$ for integers. This is the Gamma-Poisson distribution.

12.3 Negative Binomial models

Similar to the Poisson model, but allowing for more general variance functions and also more general probabilities of zero counts.

Let X be a r.v. with a negative binomial distribution with parameters p, $0 < p < 1$ and

$r > 0$. Here X can be interpreted as the number of trials it takes in a sequence of independent Bernoulli trials until one achieves r successes. The pmf of this distribution is

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x \geq r$$

with

$$E(X) = \frac{r}{p} \quad \text{and} \quad \text{var}(X) = \frac{r(1-p)}{p^2}. \tag{12.3}$$

Comparing this pmf with the pmf obtained above for Y, we find they are identical if

$$r = \kappa, \quad p = \frac{\rho}{\rho + 1} \quad \text{and} \quad X = \kappa + Y. \tag{12.4}$$

Then

$$E(Y) = E(X) - \kappa = \frac{\kappa(\rho + 1)}{\rho} - \kappa = \frac{\kappa}{\rho} = \mu, \quad \text{var}(Y) = \text{var}(X) = \kappa \, \frac{1 + \rho}{\rho^2}. \tag{12.5}$$

Note: (1) Mean and variance in (12.5) can also be derived directly by a conditioning argument, using the Poisson-Gamma representation of the negative binomial.
(2) When r is considered a nuisance parameter, the negative binomial distribution is in the one-parameter exponential family.

For large ρ and fixed μ, i.e., small variance of Z,

$$\text{var}(Y) = \frac{\kappa(1 + \rho)}{\rho^2} = \mu \, \frac{\rho + 1}{\rho} \approx \mu$$

so that Poisson regression without overdispersion will be approximately correct.

For small ρ and fixed μ, i.e., large variance of Z,

$$\mathrm{var}(Y) = \frac{\kappa(1+\rho)}{\rho^2} = \frac{\mu}{\rho}(1+\rho) \approx \frac{\mu}{\rho}.$$

Generally one will encounter overdispersion with overdispersion parameter $\sigma^2 = 1/\rho$, if a Poisson model is used. From $\mu = \frac{\kappa}{\rho}$, we find

$$\mu + \mu^2/\kappa = \kappa\,(\frac{1}{\rho} + \frac{1}{\rho^2}) = \kappa\,\frac{1+\rho}{\rho^2} = \mathrm{var}(Y).$$

Implications: The null count $P(Y=0) = (1+\rho^{-1})^{-\kappa}$ can be modeled with two parameters (enhanced flexibility) and the variance function $V(\mu) = \mu + \mu^2/\kappa$ is more flexible than the Poisson model, it allows for quadratic terms (for very large κ it is the same as Poisson). Estimation of κ not part of the GLM, it is an assumed extraneous parameter, can be estimated by method of moments (similar to the nuisance parameter, eg, σ^2 in the normal model).

12.4 Zero-inflated models

In GLMs with count data, non-zero counts may follow the assumed model, while the zero count does not (usually zeros are more frequent than predicted by the model, giving rise to an inflated probability for a zero).

Examples: (1) Counts of infected individuals in communities, to assess the spread of an epidemic, where zero counts can occur either if the epidemic has not yet reached the community (a *structural zero*), or if it has, nobody is currently infected (a *sampling zero*). This is a situation where one can expect zero inflation.

(2) Car driving accidents for adults aged between 80-90 years. Many of these seniors will

not drive anymore (structural zeros), while those who do may have zero accidents (sampling zeros) in the review period. This is a situation with likely zero inflation.

Zero-inflated models have been developed for both Poisson (ZIP) and for Negative Binomial Models (ZINB). They are mixtures of an atomic distribution with a point mass of size 1 located at 0 and Poisson (resp. Negative Binomial) distributions.

If the pmf of the Poisson or Negative Binomial distribution with mean μ is $f(y;\mu)$ with $\mu = g^{-1}(\eta) = g^{-1}(X\beta)$, then the zero-inflated model is given by

$$\begin{aligned} &P(Y_i = y|X_i) = \alpha_i + (1-\alpha_i)\, f(y;\mu_i) \quad \text{for } y = 0; \\ &P(Y_i = y|X_i) = (1-\alpha_i)\, f(y;\mu_i) \quad \text{for } y > 0; \\ &\text{logit}(\alpha_i) = \tilde{\eta}_i = X_i\gamma, \quad \text{for a parameter } \gamma \in \Re^p. \end{aligned}$$

The zero inflation may thus depend on the predictor level. Here α_i is interpreted as probability of a structural zero, in contrast to $f(0,\mu)$, which are sampling zeros. If there is no structural zero, $\alpha_i = 0$.

Fitting this model requires to jointly fit parameter vectors β and γ, by maximizing a combined likelihood. The estimated proportion of structural zeros among all zeros at covariate level X_i is

$$\frac{\alpha_i}{\alpha_i + (1-\alpha_i)\, f(0;\mu_i)},$$

which gives a measure of the degree of zero overinflation that is present in the data at predictor level x_i.

Note such mixture models can be devised for inflating other counts, for example there could be an abundance of counts of a particular size in some applications, where you only get to see pre-processed recordings where continuous outcomes are rounded always up to the next integer. In such cases one may see inflation for the outcome 1.

A related issue is *zero truncation*, where a zero in principle cannot be observed: This would be the case if one conducts a study for which a sample of patients in a hospital is recruited, where the response is the number of hospital stays during the last 10 years; obviously, the response is greater or equal than 1 and a 0 cannot occur.

A simple approach in such cases is to subtract 1 from all responses, then fitting the selected count model; a variant is that a zero outcome is possible but is under-inflated, i.e., is less likely to occur than provided for by a Poisson or other selected count model.
Then the mentioned subtraction method will lead to possible outcomes of -1: This can be handled by a mixture model where an outcome of -1 is one mixture component, occurring with probability α_i, and the second mixture component is the regular count model, occurring with probability $(1 - \alpha_i)$.

12.5 LAB 4: Overdispersion and Zero-Inflation

12.5.1 Overdispersion in Poisson Regression

Causes of over-dispersion in Poisson regression

- Clustered Poisson
- Mixture Model (heterogeneity)
- Model misspecification, important variables omitted and zero-inflation or zero-truncation could also lead to what appears an over-dispersion problem.

Approaches:

1. *Negative Binomial Regression*

   We can fit a negative binomial regression model to replace a Poisson regression model.

   ```
   library(MASS)
   ?glm.nb()
   ```

- An alternating optimization process between mean vector $\boldsymbol{\mu}$ and θ is used, until convergence occurs. For given fixed θ, the algorithm finds the optimum for $\boldsymbol{\mu}$.
- "init.theta": specifies the initial value of θ. If omitted, a moment estimator after performing an initial Poisson GLM fit is used.

```
glm(formula, family = negative.binomial(theta))
```

- If using the "glm()" function, one must specify the over-dispersion parameter θ, and the program will assume it to be known and carry out the usual IWLS to solve for the model coefficients.

Example

```
library(pscl) # containing the bioChemists data
data("bioChemists")
n <- nrow(bioChemists)
## compare Poisson and Negative Binomial Model
poisson <- glm(art ~ ., data = bioChemists, family = poisson())
sigma2 <- poisson$deviance/(n-length(coef(poisson)))
sigma2 # 1.80
# fitting nb without specifying theta
nb1 <- glm.nb(art ~ ., data=bioChemists)
summary(nb1)
nb1$deviance / (n-length(coef(nb1))) # 1.105
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.256144   0.137348   1.865 0.062191 .
femWomen    -0.216418   0.072636  -2.979 0.002887 **
marMarried   0.150489   0.082097   1.833 0.066791 .
kid5        -0.176415   0.052813  -3.340 0.000837 ***
```

```
phd          0.015271   0.035873   0.426 0.670326
ment         0.029082   0.003214   9.048  < 2e-16 ***

    Null deviance: 1109.0  on 914  degrees of freedom
Residual deviance: 1004.3  on 909  degrees of freedom
AIC: 3135.9

# fitting nb with a theta
nb2 <- glm(art ~ ., data = bioChemists, family=negative.binomial(nb1$theta))
summary(nb2)
nb2$deviance / (n-length(coef(nb2))) # 1.105

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.256149   0.140008   1.830  0.06765 .
femWomen    -0.216420   0.074043  -2.923  0.00355 **
marMarried   0.150490   0.083686   1.798  0.07247 .
kid5        -0.176416   0.053836  -3.277  0.00109 **
phd          0.015271   0.036568   0.418  0.67633
ment         0.029082   0.003277   8.876  < 2e-16 ***

    Null deviance: 1109.0  on 914  degrees of freedom
Residual deviance: 1004.3  on 909  degrees of freedom
AIC: 3133.9
# How do you explain the difference in AIC?
```

2. *Quasi-Poisson model*

   Another way of dealing with over-dispersion is to use the mean regression function and the variance function from the Poisson GLM but to leave the dispersion parameter ϕ unrestricted. Thus, ϕ is not assumed to be fixed at 1 but is estimated from the data.

This strategy leads to the same coefficient estimates as the standard Poisson model but inference is adjusted for over-dispersion.

```
glm(formula, family = quasipoisson())
```

Example

```
# Continue with previous example where we have the poisson fit
quasip <- glm(art ˜ ., data = bioChemists, family = quasipoisson())
summary(quasip)
quasip$deviance / (n-length(coef(quasip))) # 1.798
summary(quasip)$dispersion # 1.829


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.304617   0.139273   2.187 0.028983 *
femWomen    -0.224594   0.073860  -3.041 0.002427 **
marMarried   0.155243   0.083003   1.870 0.061759 .
kid5        -0.184883   0.054268  -3.407 0.000686 ***
phd          0.012823   0.035700   0.359 0.719544
ment         0.025543   0.002713   9.415  < 2e-16 ***
```

12.5.2 Zero-Inflated Poisson and Negative Binomial Model

- Fitting zero-inflated model

  Structural zeros violate the assumptions of the Poisson model and thus need to be accounted for.

  ```
  library(pscl)
  ?zeroinfl()
  ```

  – “dist”: model distribution to be specified, “poisson” (default) or “negbin”

- “link”: the link function used in the binary zero-inflation model, “logit” (default), “probit”, “cloglog”, etc.
- Specify predictors to be considered for the zero-inflation effect in the model formula (see example below for details).

Example

```
zip <- zeroinfl(art ~ . | ., data = bioChemists, dist = 'poisson')
summary(zip) # Zero-Inflated Poisson Model

Call:
zeroinfl(formula = art ~ . | ., data = bioChemists, dist = "poisson")

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.640838   0.121307   5.283 1.27e-07 ***
femWomen    -0.209145   0.063405  -3.299 0.000972 ***
marMarried   0.103751   0.071111   1.459 0.144565
kid5        -0.143320   0.047429  -3.022 0.002513 **
phd         -0.006166   0.031008  -0.199 0.842378
ment         0.018098   0.002294   7.888 3.07e-15 ***

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.577059   0.509386  -1.133  0.25728
femWomen     0.109746   0.280082   0.392  0.69518
marMarried  -0.354014   0.317611  -1.115  0.26502
kid5         0.217097   0.196482   1.105  0.26919
phd          0.001274   0.145263   0.009  0.99300
ment        -0.134114   0.045243  -2.964  0.00303 **
```

```
zinb <- zeroinfl(art ˜ . | ., data = bioChemists, dist = ’negbin’)
summary(zinb) # Zero-Inflated Negative Binomial Model

Call:
zeroinfl(formula = art ˜ . | ., data = bioChemists, dist = "negbin")

Count model coefficients (negbin with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.4167465  0.1435966   2.902  0.00371 **
femWomen    -0.1955068  0.0755926  -2.586  0.00970 **
marMarried   0.0975826  0.0844520   1.155  0.24789
kid5        -0.1517325  0.0542061  -2.799  0.00512 **
phd         -0.0007001  0.0362697  -0.019  0.98460
ment         0.0247862  0.0034927   7.097 1.28e-12 ***
Log(theta)   0.9763565  0.1354695   7.207 5.71e-13 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.19169    1.32282  -0.145  0.88478
femWomen     0.63593    0.84892   0.749  0.45379
marMarried  -1.49947    0.93867  -1.597  0.11017
kid5         0.62843    0.44278   1.419  0.15582
phd         -0.03771    0.30801  -0.122  0.90254
ment        -0.88229    0.31623  -2.790  0.00527 **
```

- Test for existence of zero-inflation in R

```
libary(pscl)
?vuong(model1, model2)
```

The Vuong non-nested test is based on a comparison of the predicted probabilities of

two models that are not nested. Examples include comparisons of zero-inflated count models with their non-zero-inflated analogs. In R, the two model objects can inherit from “glm”, “negbin” or “zeroinfl” objects. Essentially, we do a hypothesis test with:
H_0: the two models are indistinguishable,
H_1: the two models are different.
The test statistic is asymptotically standard Normal, and we reject the null whenever it is large in magnitude.

Example

```
vuong(poisson, zip) # Poisson vs. zero-inflated Poisson

Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishable)
-------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                   -4.180493 model2 > model1 1.4544e-05
AIC-corrected         -4.179901 model2 > model1 1.4582e-05
BIC-corrected         -2.332762 model2 > model1  0.0098303

vuong(nb1, zinb) # NegBin vs. zero-inflated NegBin

Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishable)
-------------------------------------------------------------
              Vuong z-statistic             H_A  p-value
Raw                   -2.241831 model2 > model1 0.012486
AIC-corrected         -2.240491 model2 > model1 0.012530
BIC-corrected          1.939690 model1 > model2 0.026209
```

13 Bootstrap inference in the GLM

13.1 Residual Bootstrap

Bootstrap is a tool to obtain asymptotic valid inference through resampling with replacement from the available data. The resampling procedure introduces randomness which then gives rise to a CLT that applies under regularity conditions and justifies the bootstrap.

Given independent data (X_i, Y_i), $i = 1, \ldots, n$, with predictors X_i and responses Y_i, we fit a regression model $E(Y|X = x) = f(x, \beta)$ for a parametric regression model $f(\cdot, \cdot)$ with regression parameter vector β.

Examples: Multiple linear regression $f(x, \beta) = x^T\beta$
GLM $f(x, \beta) = g^{-1}(x^T\beta)$ for a link function g.

The fitted model is $f(x, \hat{\beta})$, where $\hat{\beta}$ is a suitable estimate of β, most commonly obtained by least squares (in a multiple linear model) or by MLE.

There are two regression bootstraps, the *residual bootstrap* described in the following and the *random X bootstrap* described in the next section.

Obtain residuals $r_i = Y_i - f(X_i, \hat{\beta})$ and form the sample of residuals $R_n = (r_1, \ldots, r_n)$. Then B times obtain a sample of n new residuals by resampling from R_n with replacement, i.e., each time a bootstrap sample is obtained, any of the r_i is selected with probability $1/n$, until one has n elements in the bootstrap sample. These samples form the bootstrap residual samples R_b^*, $b = 1, \ldots B$, each consisting of n elements, $R_b^* = (r_{1b}^*, \ldots, r_{nb}^*)$. Here B is typically chosen as $B = 2000$ or larger.
Then form the b-th bootstrap data sample $S_b^* = \{(X_{1b}^*, Y_{1b}^*), \ldots, (X_{nb}^*, Y_{nb}^*)\}$, where $X_{kb}^* = X_k$, $k = 1, \ldots, n$, $b = 1, \ldots B$, and $Y_{kb}^* = f(X_{kb}^*, \hat{\beta}) + r_{kb}^*$, $k = 1, \ldots, n$, $b = 1, \ldots B$.
We obtain the b-th bootstrap estimate $\hat{\beta}_b^*$, $b = 1, \ldots, B$, by applying the chosen estimation

method to the b-th bootstrap sample S_b^*.

Note: Since we add residuals to the fitted model, this residual bootstrap only works for a homoscedastic regression model, where the error distribution does not depend on the predictors. This means it cannot be used for GLM. It also does not reflect variation in the predictor levels which are not resampled, therefore it is best suited for situations where the predictors are non-random.

13.2 Random X bootstrap

Assuming a random design, taking into account the randomness of both X and Y, one resamples with replacement directly from the originally observed sample (X_i, Y_i), $i = 1, \ldots, n$. Doing this B times yields the bootstrap samples. This method takes the random variation of X into account and makes no assumptions about the dependence of the distribution of the residuals on the predictor.

Then the b-th bootstrap data sample is $S_b^* = \{(X_{1b}^*, Y_{1b}^*), \ldots, (X_{nb}^*, Y_{nb}^*)\}$, where X_{kb}^*, Y_{kb}^*, $k = 1, \ldots, n$, is obtained by resampling from the original sample with replacement. As in the residual bootstrap method, we obtain the b-th bootstrap estimate $\hat{\beta}_b^*$, $b = 1, \ldots, B$, by applying the chosen estimation method to the b-th bootstrap sample S_b^*.

For both residual and random X bootstrap, we obtain confidence regions from the bootstrap sample of the parameter estimates $\hat{\beta}_b^*$, $b = 1, \ldots, B$.

For one component β_p of the parameter vector β, the $(1-\alpha)$ c.i. $[L, U]$ is obtained from the empirical distribution function F_p of the r.v.s $\hat{\beta}_{pb}^*$, $b = 1, \ldots, B$. Denoting the corresponding quantile function by Q_p, we set $L = Q_p(\alpha/2)$, $U = Q_p(1-\alpha/2)$.

For simultaneous $(1-\alpha)$ confidence regions for β or a subset of β we can fit a multivariate normal distribution to the random vectors $\hat{\beta}_b^*$ and then determine the $(1-\alpha)$ contour ellipsoid; approximately a fraction α of the $\hat{\beta}_b^*$ will fall outside this ellipsoid and a fraction $(1-\alpha)$ inside. We can also obtain bootstrap hypothesis tests (see the following Lab section for an example).

13.3 LAB 5: Bootstrap

Random X Bootstrap

We draw bootstrap samples with replacement from the data (X_i, Y_i), $i = 1, \ldots, n$. We can then construct bootstrap confidence intervals and perform hypothesis tests for the regression coefficients. A convenient tool in R to implement bootstrap is the `boot` function in the `boot` package.

- Three important arguments of `boot`:
  - `data`: A data vector, matrix, or data frame holding the original data to be sampled from.
  - `statistic`: A function that returns the statistic of interest. The first argument of this function will always be the original data. The second will be a vector of indices, frequencies or weights which define the bootstrap sample.
  - `R`: The number of bootstrap replications, denoted by B in Sections 13.1–13.2.
- `dataEllipse`: A function to construct confidence region with the `car` package.

Example 1: Bootstrap Confidence Region for β

For melanoma data, consider

$$\log(E(\text{total incidence}|X)) = \beta_0 + \beta_1(\text{year}) + \beta_2(\text{number of sunspot}).$$

- Bootstrap confidence interval for β_j is $[L_j, U_j]$, where L_j and U_j are the $[(B+1)\alpha/2]$-th and $[(B+1)(1-\alpha/2)]$-th smallest values in the bootstrap samples of estimates $\{\hat{\beta}_{jb} : b = 1, \ldots, B\}$. Here, $[a]$ denotes the nearest integer to a.
- Simultaneous confidence regions/ellipses for $(\beta_1, \beta_2)^\top$: Use `dataEllipse()`. 95% and 99% bootstrap confidence regions for $(\beta_1, \beta_2)^\top$ are shown in Figure 2.

```
B <- 999
data <- read.table('melanoma.txt', header=TRUE)
beta.est <- function(data, ind){
res <- glm(totalincidence ~ ., data=data[ind,], family=poisson())
coef(res)
}
boot.res <- boot(data, beta.est, R=B)

alpha = 0.05
# 95% bootstrap confidence interval for beta_1
sort(boot.res$t[,2])[round(c((B+1)*alpha/2, (B+1)*(1-alpha/2)))]
# 95% bootstrap confidence interval for beta_2
sort(boot.res$t[,3])[round(c((B+1)*alpha/2, (B+1)*(1-alpha/2)))]

library(car)
# 95% and 99% bootstrap confidence regions for (beta_1, beta_2)
dataEllipse(boot.res$t[,2], boot.res$t[,3],
xlab= 'year coefficient', ylab='sunspot coefficient',
cex=.3, levels=c(.95, .99))
```

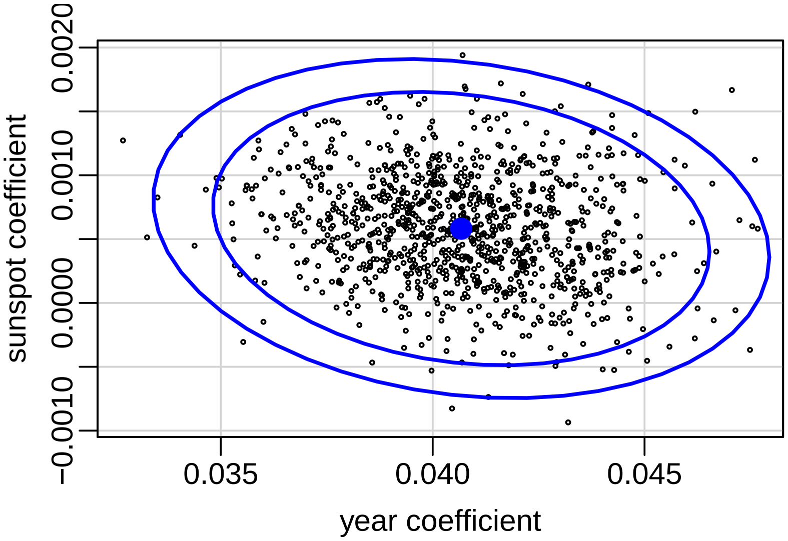


Figure 2: Bootstrap confidence regions for $(\beta_1, \beta_2)^\top$.

Example 2: Bootstrap Hypothesis Testing

Following Example 1. Suppose we want to test

$$H_0: \ \beta_1 = \beta_2 = 0$$

$$H_a: \ \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

- Calculate the estimates for the coefficient vector $\hat{\beta}$ and asymptotic covariance matrix $\hat{\Sigma}$ and the test statistic $T = \hat{\beta}^\top \hat{\Sigma}^{-1} \hat{\beta}$ based on the original data.
- Calculate the test statistic $T_b^* = (\hat{\beta}_b^* - \hat{\beta})^\top \hat{\Sigma}^{-1} (\hat{\beta}_b^* - \hat{\beta})$ for each bootstrap sample b, $b = 1, \ldots, R$.
- Calculate the bootstrap p-value: $(\#\{T_b^* > T : b = 1, \ldots, R\} + 1)/(R + 1)$.

```
res <- glm(totalincidence ~ ., data=data, family=poisson())
invSigma <- solve(vcov(res)[2:3,2:3])
beta <- coef(res)[2:3]


beta.test <- function(data, ind){
res <- glm(totalincidence ~ ., data=data[ind,], family=poisson())
beta.star <- coef(res)[2:3]
t(beta.star-beta) %*% invSigma %*% (beta.star-beta)
}
boot.test <- boot(data, beta.test, R=B)


## bootstrap p-value
(sum(as.vector(boot.test$t) > as.vector(t(beta) %*% invSigma %*% beta)) + 1) /
(B + 1)
```

13.4 PROBLEM SET 4

1. Show that a negative binomial model belongs to the exponential family if the number of successes r is considered known.

2. You are asked to model the number of epileptic seizures of patients within a fixed time period, say a month. Assume that the counts Y of seizures are Poisson distributed, $Y \sim$ Poisson(Z), where Z itself is a patient-dependent random variable that may be assumed to be Gamma-distributed, $Z \sim$ Gamma(κ, ρ). It is known that

$$E(Z) = \frac{\kappa}{\rho}, \quad \text{var}(Z) = \frac{\kappa}{\rho^2}.$$

Use a conditioning argument (note: do not use the Gamma density) to show that

(a)
$$E(Y) = \frac{\kappa}{\rho}$$

(b)
$$\text{var}(Y) = \frac{\kappa(\rho+1)}{\rho^2}.$$

(c) Write the variance as a function of μ and ρ and investigate its behavior for the case of small and large ρ.

3. For the Aranda-Ordaz link transformation family, derive the form of the additional predictor γ_i in (10.5) for general α and for the special case $\alpha_0 = 1$.

4. Derive (11.4) and (11.5). Derive the fact that the baseline odds model (11.6) can be equivalently represented as a series of logistic regression models as per (11.7).

5. Derive the direct representations for π_{ij} at the end of section 11.2.

6. For the epileptic seizure data in chemo.dat, using Poisson regression with log link,

    (a) Obtain and interpret Cook's distance and leverage plots.

    (b) Check for overdispersion. How might this arise?

    (c) Fit a negative binomial model to these data. Check goodness of fit. Do your results change?

    (d) Fit a quasi-likelihood model with log link function. Is there any difference to the previous analysis?

7. In an epidemiological field study, a researcher wishes to quantify the occurrence of schistosomiasis, a parasitic disease in the tropics incurred from freshwater snails. For this purpose, the researcher travels to remote villages in order to count the number of infected individuals, and also to assess a snail control index, 0=no control, 1=good control, a hygiene index, 0=bad, 1=medium, 2=good and a medical access index, indicating access to a medical facility, and coded as 0=bad, 1=good. A fourth predictor is the number of inhabitants of the village. The researcher asks you to analyze these data.

    (a) Write a generalized linear model that is appropriate as an initial approach for the analysis. Provide all components of the model and state the needed assumptions. How is the hygiene index coded and what would be the natural baseline level?

    (b) Discuss the role of the predictor no. of inhabitants of the village. Do you expect this to be a relevant predictor? If you find it to be significant, how do you proceed with the remaining analysis?

    (c) Sketch the construction of a 95% confidence ellipsoid for the two parameters associated with the snail control index and the medical access index. Please indicate which quantities in the output after fitting the model you would use for this.

    (d) The researcher would like to find out whether one may conclude from these data that controlling snails is more effective than improving medical access. In the model you have developed, find a suitable null hypothesis and alternative to establish this, in a way that will provide as much power as possible for drawing the conclusion that the researcher is interested in. Indicate in which way this could be tested in the framework of a GLM and provide the distribution of the test statistic under the null hypothesis.

    (e) Assume for the following that there is a sizable number of villages where nobody is at risk for catching the disease, due to the fact that there is no water nearby which houses the snails, but that this fact is unknown to the researcher. Write all components of a model that appropriately reflects this fact.

8. The wine data have been collected for $n = 1599$ red wines. Predictors include 11 chemical measurements made for each wine, as follows:

   (1) fixed acidity (2) volatile acidity (3) citric acid (4) residual sugar (5) chlorides (6) free sulfur dioxide (7) total sulfur dioxide (8) density (9) pH (10) sulphates (11) alcohol contents

   The response is “quality” (recorded as a score between 1 and 6), based on sensory data (median of at least 3 evaluations made by wine experts). Thus every row in the file has 12 variables and there are 1599 rows.

   (a) Please find a suitable model for the dependence of wine quality on the chemical ingredients. Merge scores 1-3 into a low quality category, consider 4 the medium category and merge 5-6 to form the good quality category, for a total of 3 categories. Compare the fitting of the proportional odds and baseline odds models. Which works better for these data? Which of the predictors determine the quality of the wine? Discuss your findings.

   (b) Now form just two categories, 1-3 and 4-6, to distinguish the good and not so good wines. For this new type of response, fit a logistic regression model. Determine which predictors are relevant for the quality of the wine in this simplified model.

   (c) Compare the fits and inference of the binomial regression with those of the multinomial regression models. Which modeling approach do you prefer for this application? Give reasons.

9. For the simplified responses as in the previous question, fit a GAM and report all relevant outputs. Interpret the additive functions for the predictors that are included after predictor selection. Provide interpretations.

10. Show for the hat matrix H that (denoting its diagonal elements by h_{ii}) $\sum_{i=1}^{n} h_{ii} = p$.

11. In a study in neuroscience, one measures the numbers of spikes elicited in a specific brain cell during an eye movement. The number of spikes s is recorded in a 50ms time interval and is a response to an experimental stimulus in which the head of the subject is turned with a specific speed v and acceleration a, both of which are recorded, along with an eye movement velocity e. This experiment is repeatedly carried out for $n = 50$ sessions, where we may assume that the recordings are independent for different sessions.

    (a) Provide the components of a GLM, by which the response can be related to the predictors. Choose a suitable link function.

(b) It is found that some cells are in an inactive state where they cannot respond to the stimulus and therefore no spikes are recorded. It is unknown which of the cells are in the inactive state. Propose an approach to handle this situation and provide detailed descriptions of all components.

(c) It is suggested that an interaction effect may exist between e and a and between e and v. Please describe how you would extend GLM to include such an effect.

(d) Describe the considerations you would use to decide whether to use a Poisson model, zero inflated Poisson model, negative binomial model, or zero inflated negative binomial model when analyzing these data.

12. Assume that in the previous problem there are no cells in the inactive state. However, due to a recording machine error the spike recordings of a fraction of all outcomes are changed to an outcome of 10 spikes.

    (a) Provide a model for this situation.

    (b) Describe how you fit the model to the data.

    (c) Give an estimator for the fraction of recordings that is affected.

13. Implement the random X bootstrap for the melanoma data. Specifically, use the bootstrap for problems 6b and 7 in Set 1 and compare with the previous solutions.

14. Implement the random X bootstrap for the Pima Indians data. Specifically, use the bootstrap to construct 95% a c.i. for each of the slope parameters and compare with the c.i.s obtained from the asymptotic normality of the MLEs.

14 Constant coefficient of variation models

14.1 Gamma regression

Coefficient of variation:

$$\mathrm{v} = \frac{\mathrm{var}(y)}{(Ey)^2}$$

Data with constant coefficient of variation satisfy $\sigma \sim Ey$, and

$$\mathrm{var}(y) = \mathrm{v}\,(Ey)^2 = \mathrm{v}\mu^2, \;\; \mu = Ey$$

Note: Constant coefficient of variation data are always non-negative. Then

$$\log y = \log(Ey) + (y - Ey)\tfrac{1}{Ey} - \tfrac{1}{2}(y-Ey)^2\tfrac{1}{(Ey)^2} + R \quad \text{for a remainder } R$$

$$\Rightarrow E\,(\log y) \approx \log\mu - \frac{1}{2}\mathrm{v},\ \ \mathrm{var}(\log y) \approx \frac{\mathrm{var}(y)}{\mu^2} = \mathrm{v} \equiv \ \text{const},$$

assuming remainder R is negligible. Therefore the log transformation stabilizes the variance for data with constant coefficient of variation.

Conclusion: Can run ordinary LSE's after log-transformation; the intercepts will be biased by the *off-set* $-\frac{1}{2}\mathrm{v}$.

Pdf for Gamma data:

$$f_{\mu,\nu}(y) = \frac{1}{\Gamma(\nu)}\,(\frac{\nu y}{\mu})^\nu \,\exp\,(-\frac{\nu y}{\mu})\,\frac{1}{y},\ \ y \geq 0 \tag{14.1}$$

Note: Relating (12.1), (14.1),

$$\kappa = \nu,\ \ \rho = \frac{\nu}{\mu};\quad Ey = \mu,\ \ \mathrm{var}(y) = \frac{\mu^2}{\nu},\ \ \mathrm{v} = \frac{1}{\nu}$$

$\Rightarrow$ Gamma data ($\nu \equiv$ const) have constant coefficient of variation $\mathrm{v} = \frac{1}{\nu}$.

Canonical link is inverse function $g(x) = \frac{1}{x}$, often not very practical; log link is more appropriate in most cases.

14.2 Transformation models

Box-Cox transformations: $h(x) = x^\alpha$, $x \geq 0$, $\alpha > 0$ or $h(x) = \log(x)$, $x > 0$.

We can transform predictors or responses. Transforming predictors is the first thing to do when there is lack of fit in a GLM.

Linear predictor with transformed variables: $\eta = \sum_{j=1}^{q}\beta_j h_j(x_j)$, where h_j in the simplest case could be linear and squared predictors (if continuous) or can include Box-Cox transformed predictors; usually will keep original predictors and add transformed variables or squares as additional predictors → $q > p$, adding too many transformed predictors can be a concern. Also can add *interaction terms*, usually just products between predictors and usually only if there is a substantial justification.

Transformation Models: Transform responses with transformation h, usually $h : \mathcal{R}^+ \to \mathcal{R}$ to obtain the model $E(h(Y)|X) = \eta$, or $h(Y) = X\beta + e$ for additive errors e with mean zero and finite variance that are independent of X.

Example: *Log-additive model*

$$\log y = \mu + \epsilon, \;\; E\epsilon = 0, \;\; \mathrm{var}(\epsilon) = \sigma^2$$

$$Ey = e^{\mu}\, E\, e^{\epsilon} \approx e^{\mu} =: \tilde{\mu}, \quad \mathrm{var}(y) \approx (e^{\mu})^2 \mathrm{var}(e^{\epsilon}) \approx \tilde{\mu}^2\sigma^2.$$

This model behaves very similarly to a constant-coefficient of variation model and but is simpler to implement, much simpler than Gamma regression. Often a good alternative to Gamma regression.

Lognormal model: $\epsilon \sim N(0, \sigma^2)$. Provides a simple approach that often works sufficiently well for constant coefficient of variation situations is to apply a log-transformation of y followed by an ordinary multiple linear regression fit.

Note: Transformation models are not GLMs – why? In transformation models, the effects are fitted at the transformed level. Since the transformation is generally nonlinear, the interpretation after the inverse transformation is applied to the fitted model in order to describe the findings on the original scale can be tricky.

15 Quasi-Likelihood

15.1 Quasi-Scores

Motivation: Most features of GLMs depend on the first two moments only, rather than the entire distribution. Mean function, $\mu = g^{-1}(\eta)$, and variance function $V(\mu)$ determine the model.
Hence, rather than specifying the exponential family distribution, it is sufficient to specify:

1. *Link function* defining $\mu = g^{-1}(\eta), \ \ Ey = \mu$

2. *Variance function* $V(\mu)$, $\ \ \text{var}(y) = \sigma^2 V(\mu)$. (Earlier a derived, now a pre-specified model component).

3. Independence of observations

Under these minimal assumptions, there is no likelihood function–we do not have pdf's.

Quasi-score as alternative:

$$U = u(\mu; y) = \frac{y - \mu}{\sigma^2 V(\mu)}$$

Score is first considered as a function of μ, later as a function of the parameters β.
Observe: $\ EU = 0, \quad \text{cov}(U) = (\sigma^2 V(\mu))^{-1},$

$$E(\frac{\partial U}{\partial \mu}) = E(\frac{-\sigma^2 V(\mu) - (y-\mu)\sigma^2 V'(\mu)}{(\sigma^2 V(\mu))^2}) = -(\sigma^2 V(\mu))^{-1}.$$

So we have

$$EU = 0, \quad \text{cov}(U) = -E(\frac{\partial U}{\partial \mu}), \ \ \text{cov}(U) + E(\frac{\partial U}{\partial \mu}) = 0,$$

meaning that the first two Bartlett identities are satisfied for the Quasi-score.

15.2 Quasi-information and Quasi-likelihood

This motivates to define:

Quasi-Information (expected):

$$i_\beta = -E\,(\frac{\partial U}{\partial \beta})$$

Quasi-Likelihood (QL) for one observation:

$$Q(\mu, y) = \int_y^\mu \frac{y-t}{\sigma^2 V(t)}\, dt,$$

noting

$$U = \frac{\partial Q}{\partial \mu} = \frac{y-\mu}{\sigma^2 V(\mu)};$$

QL for all observations:

$$Q(\mu, y) = \sum_{i=1}^n Q_i(\mu_i, y_i) = \sum_{i=1}^n \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2 V(t)}\, dt$$

plays the role of log-likelihood $l = \log L$.

Since $Q(y, y) = 0$, we obtain *Quasi-deviance*

$$D(y, \mu) = 2\sigma^2\{Q(y, y) - Q(\mu, y)\} = 2\sum_{i=1}^n \int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)}\, dt > 0.$$

Each summand is non-negative (for $y_i \geq \mu_i$ as well as $y_i \leq \mu_i$).

Examples

(1) $V(\mu) = 1,\ \sigma^2 < \infty :$ ”Quasi-normal”

$$Q(\mu, y) = \int_y^\mu \frac{y-t}{\sigma^2}\, dt = \frac{1}{\sigma^2}\int_0^{\mu-y} (-v)\, dv = -\frac{1}{2\sigma^2}(y-\mu)^2$$

(2) $V(\mu) = \mu(1-\mu),\;\; \sigma^2 = 1,\;\; y = \begin{cases} 0 \\ & \text{binary} \\ 1 \end{cases}$

Use $\frac{d}{dt}\log(\frac{t}{1-t}) = \frac{1}{t(1-t)},\;\; \frac{d}{dt}\log(1-t) = -\frac{1}{1-t}$ to obtain

$$\begin{aligned} Q(\mu, y) &= \int_y^\mu \frac{y-t}{t(1-t)}\,dt = y\,\log(\frac{t}{1-t})|_y^\mu + \log(1-t)|_y^\mu \\ &= \begin{cases} \log(1-\mu) & y = 0 \\ \\ \log\mu & y = 1 \end{cases}\} = y\log(\frac{\mu}{1-\mu}) + \log(1-\mu) \\ &= y\,\text{logit}(\mu) \,+\, \log(1-\mu), \end{aligned}$$

the same as the log-likelihood l for the binomial case: "Quasi-binomial" case.

15.3 Estimating equations

Quasi-likelihood in the parameters β:

$$\mu = \mu(\beta) = g^{-1}(\beta^\top x)$$

$$Q(\beta, y) = \int_y^{\mu(\beta)} \frac{y-t}{\sigma^2 V(t)}\,dt$$

$$\frac{\partial}{\partial\beta_j}Q(\beta, y) = U_j(\beta) = \frac{y-\mu}{\sigma^2 V(\mu)}\frac{\partial\mu}{\partial\beta_j} \quad \text{(chain rule)}$$

$$U_j(\beta) = \sum_{i=1}^n \frac{\partial\mu_i}{\partial\beta_j}\,\frac{y_i-\mu_i}{\sigma^2 V(\mu_i)}, \qquad 1 \le j \le p$$

Define $V = diag(V(\mu_1), \ldots, V(\mu_n)),\;\; n \times n$

$D = (d_{ij}) = (\frac{\partial\mu_i}{\partial\beta_j})_{1\le i\le n, 1\le j\le p} \qquad n \times p$

$$\Rightarrow U(\beta) = \begin{pmatrix} U_1(\beta) \\ \vdots \\ U_p(\beta) \end{pmatrix} = D^\top V^{-1}(y-\mu)/\sigma^2,$$

therefore

$$\begin{aligned} E\,U(\beta) &= 0, \\ \mathrm{cov}(U(\beta)) &= \frac{1}{\sigma^4}(D^\top V^{-1})\mathrm{cov}(y)(D^\top V^{-1})^\top \\ &= D^\top V^{-1} D/\sigma^2 \quad \text{ since } \mathrm{cov}(y) = \sigma^2 V. \end{aligned}$$

Observe for the quasi-information matrix in β, corresponding to Fisher information in the GLM, that

$$i_\beta = -E(\frac{\partial U}{\partial \beta}) = X^\top W X = D^\top V^{-1} D/\sigma^2 = \mathrm{cov}(U(\beta)), \tag{15.1}$$

which follows by direct calculation, using $Ey = \mu$ and similar arguments as in (8.11).

QL equations or *estimating equations:*

$$U(\beta) = 0$$

Iterative updating analogous to Fisher scoring in regular likelihood, as quasi-information corresponds to expected information in regular likelihood:

$$\hat\beta_{(l+1)} = \hat\beta_{(l)} + i^{-1}_{\hat\beta_{(l)}} U(\hat\beta_{(l)}) = \hat\beta_{(l)} + [D^\top(\hat\beta_{(l)}) V^{-1}(\hat\beta_{(l)}) D(\hat\beta_{(l)})]^{-1} [D^\top(\hat\beta_{(l)}) V^{-1}(\hat\beta_{(l)})](y - \hat\mu(\hat\beta_{(l)}))$$

One-step analysis:

$$\hat\beta = \beta + (D^\top V^{-1} D)^{-1} D^\top V^{-1}(y - \mu),$$

which is linear in y, where β is true parameter vector and D, V are the true matrices.
This means

$$\hat\beta = \beta_0 + \sum w_i y_i$$

so that a CLT can be applied to infer asymptotic normality.

Note

$$\text{cov}(\hat{\beta}) = (D^\top V^{-1} D)^{-1} D^\top V^{-1} \sigma^2 V V^{-1} D (D^\top V^{-1} D)^{-1} = \sigma^2 (D^\top V^{-1} D)^{-1} = i_\beta^{-1}.$$

Under regularity assumptions, including

$$(\frac{D^\top V^{-1} D}{n})^{-1} \rightarrow_{n\rightarrow\infty} \Sigma,$$

one obtains

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_D N(0, \sigma^2 \Sigma).$$

The finite quasi-information matrix is

$$i_\beta = D^\top V^{-1} D / \sigma^2,$$

whereas the asymptotic quasi-information corresponds to the limit

$$i_{\beta,\infty} = \lim_{n\rightarrow\infty} \frac{1}{n} i_\beta = (\sigma^2 \Sigma)^{-1}.$$

As before we use the moment estimators

$$\hat{\sigma}^2 = \frac{1}{n-p} P = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \approx \frac{1}{n-p} D(\hat{\mu}, y).$$

Usually, quasi-likelihood gives the same asymptotics as GLM under weaker assumptions (do not need exponential family). The QL approach requires only the choice of link and variance functions. Newton-Raphson will always use quasi-information, which corresponds to Fisher information and is non-random.

16 Further Extensions of the GLM

16.1 Single and Multiple Index Models

Index: a linear combination of the predictor variables. Models with M indices include linear predictors

$$\eta_{ij} = \sum_{k=1}^{p} x_{ik}\beta_{jk},\ i = 1, \ldots, n,\ j = 1, \ldots, M.$$

The M indices can be viewed as projections of the data on the vectors $\beta_j = (\beta_{j1}, \ldots, \beta_{jp})^\top$. These models may or may not include a variance function.

Generally one assumes that

$$E(Y|X) = h(\eta_1, \ldots, \eta_M)$$

for a multiple index model with a smooth link function $h : \mathcal{R}^M \to \mathcal{R}$, that may be assumed known or unknown. *Single index models* correspond to the case $M = 1$.

Examples: GLM, Projection Pursuit. Difference to GLM is that no relation between mean and variance is assumed and Y is not in a specified class of distributions like in GLM. Single and multiple index models are popular in social sciences and econometrics.

16.2 Generalized Additive Models (GAM)

Basic idea is to consider extensions of linear models to their additive counterparts:

1. The linear model

   $$E(Y|X = x) = \alpha + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

   can be extended to the *additive model*

   $$E(Y|X = x) = \alpha + f_1(x_1) + \ldots + f_{p-1}(x_{p-1})$$

2. The generalized linear model with link function g

   $$g(E(Y|X = x)) = \alpha + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

can be extended to the *generalized additive model (GAM)*

$$g(E(Y|X = x)) = \alpha + f_1(x_1) + \ldots + f_{p-1}(x_{p-1}).$$

For identifiability, we require $E(f_j(X_j)) = 0$, $j = 1, \ldots, p - 1$.

Further extensions:

1. *Additive model with interactions.* Introducing two-way interactions by adding smoothing surfaces (also higher-order interactions):

$$g(E(Y|X = x)) = \alpha + \sum_{j=1}^{p-1} f_j(x_j) + \sum_{j,k=1,j<k}^{p-1} f_{jk}(x_j, x_k),$$

   where the f_{jk}, $j < k$, are functions $\mathcal{R}^2 \to \mathcal{R}$ with $E(f_{jk}(X_j, X_k)) = 0$.

2. *Generalized Additive Partial Linear Model.* Assume in addition to predictors $X_1, \ldots, X_{p-1}$, we have another set of covariates $Z_1, \ldots Z_q$. Then assume

$$g(E(Y|X = x, Z = z)) = \alpha + \sum_{j=1}^{p-1} f_j(x_j) + \sum_{k=1}^{q} \beta_k Z_k,$$

   i.e., predictors Z have a linear and predictors X a nonlinear additive effect. Predictors that are categorical will be usually selected to have a linear effect, as additive modeling is not feasible for such predictors.

3. *Generalized Partial Linear Model.* Again, have in addition to predictors $X_1, \ldots, X_{p-1}$ another set of covariates $Z_1, \ldots Z_q$. Assume

$$g(E(Y|X = x, Z = z)) = \alpha + f(x_1, \ldots, x_{p-1}) + \sum_{k=1}^{q} \beta_k Z_k,$$

   i.e., predictors X have a joint nonlinear effect, where f is a "smooth" function. This can work for $p = 2, 3$, but typically not for higher dimensions, due to the curse of dimensionality.

Why not extend directly to a multivariate nonparametric regression

$$E(Y|X = x) = f(x_1, \ldots, x_{p-1})$$

for a smooth function f?

This runs into the *curse of dimensionality* for $p - 1 > 3$: Optimal MSE rate of convergence cannot be better than $n^{-4/(4+(p-1))}$, if function f is twice differentiable (the usual assumption). In one-dim case $p-1 = 1$, rate is $n^{-4/5}$ while parametric rate is n^{-1}. For large enough p, the rate becomes slower than any rate $n^{-\varepsilon}$, $\varepsilon > 0$.

Theoretical support for additive models is provided by the fact that the additive structure retains the one-dimensional rate $n^{-4/5}$ for the nonparametric components and therefore is not subject to the curse of dimensionality.

16.3 Implementation of GAM

Key steps in the implementation of GAM:

(A) Through the IWLS equivalence with Fisher scoring as shown before, the algorithm for fitting this model is reduced to fitting an additive model by weighted least squares. For this step there are several options:

(a) Obtain $Z, Z' = W^{1/2}Z$ and $X' = W^{1/2}X$ (see section 9) and regress Z' on X' by fitting the *additive model*

$$E(Z_i'|X_i' = x_i') = \alpha' + \sum_{j=1}^{p-1} f_j'(x_{ij}'). \tag{16.1}$$

Then transform back to obtain

$$f_j(x_{ij}) = \frac{1}{w_{ii}^{1/2}} f_j'(x_{ij} w_{ii}^{1/2}),$$

noting

$$E(Z_i|X_i = x_i) = \frac{1}{w_{ii}^{1/2}} E(Z_i'|X_i' = x_i w_{ii}^{1/2}),$$

for $W = \text{diag}(w_{11}, \ldots, w_{nn})$. Interpolate to obtain $f_j(u)$ on an interval.

(b) Fit

$$E(Z_i|X_i = x_i) = \alpha + \sum_{j=1}^{p-1} f_j(x_{ij})$$

directly, using case weights w_{ii} in the smoothing steps (applicable for all smoothing estimators that can be written in the form of splines or local least squares; see R routine gam).

(B) For fitting (16.1), as required for each iteration step of IWLS, rewriting f_j', x_{ij}' by f_j, x_{ij}, one has several options:

(a) B- or P-splines with suitable penalty parameters for fitting the f_j (R mgcv).

- Write $f_j(x_j) = \sum_{k=1}^{K} \beta_{jk} B_k(x_j)$, for a suitable $K = m + 3$, where m knots are used (typically equidistantly spaced) and the B_k are cubic B-spline basis functions with coefficients β_{jk}.

- Use linear predictors $\eta = \beta_0 + \sum_{j=1}^{p-1} \sum_{k=1}^{K} \beta_{jk} B_k(x_j)$ and implement GLM fit as before with IWLS.

- This can be described as flexible parametric fitting

(b) *Backfitting or Gauss-Seidel iteration.* This is an iterative algorithm (R gam):

- Initialize $\alpha^{(0)} = \hat{\beta}_j, \quad f_j^{(0)}(x_j) = \hat{\beta}_j x_j$ for initial multiple least squares parameter estimates $\hat{\beta}$, $j = 1, \ldots, p-1$.

- Update by cycling through $j = 1, \ldots, p$, using notation S for a smoother (section 3.1),

$$f_j^{(m+1)}(x) = S(x; (x_{ij},\, y_i - \alpha^{(m)} - \sum_{k=1,k\neq j}^{p-1} f_k^{(m)}(x_{ik}))_{i=1,\ldots,n}, h) \tag{16.2}$$

$$\alpha^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p-1} f_j^{(m)}(x_{ij})) \tag{16.3}$$

  where h is a smoothing parameter.

- Iterate until convergence.

(c) *Smooth Backfitting.* Efficient version but complex algorithm, based on kernel smoothing.

Output of interest: Component functions f_j and inference with p-values for components and confidence intervals for function estimates f_j.

16.4 Model comparison criteria

Often, the need arises to compare various model fits.

Model diagnostics for goodness-of-fit assessment:

- Check Residuals vs Predicted

- Runs test (sometimes of limited value in binomial regression – it tends to give rise to false alarms)

- Leverage and Cook's distance.

Actions:

- Change model, variance function or link function

- Change linear predictor, typically adding quadratic components

- Use GAM or GPLAM to determine the effect of continuous predictors

Prediction criteria: Cross-validated (CV) squared prediction error (SPE):

$$SPE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i^{(-i)})^2 \tag{16.4}$$

Alternative: k-fold CV: Remove $[\frac{n}{k}]$ subjects at a time from the sample, and predict their outcomes (large gains in computing). Do this k times for disjoint sets of leave-out data.

Fraction of variance explained:

An informal criterion which is easy to understand and communicate; also known as "quasi-R^2":

$$\hat{R}_Q^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \tag{16.5}$$

provides "fraction of variance of the response explained by the model".

Modeled after the coefficient of determination in the linear model,

$$\hat{R}^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \to R^2 = \frac{\mathrm{var}(E(Y|X))}{\mathrm{var}(Y)} = 1 - \frac{E(\mathrm{var}(Y|X))}{\mathrm{var}(Y)}.$$

16.5 LAB 6: Generalized Additive Models in R

A generalized additive model (GAM) is a generalized linear model where the linear predictor is given by a user specified sum of smooth functions of the covariates plus a conventional parametric component of the linear predictor. An example of Poisson regression model with additive predictors is

$$\log(E(Y|X_1, X_2)) = \alpha + f_1(X_1) + f_2(X_2),$$

where α is the intercept term capturing EY, and f_1 and f_2 are smooth functions of predictors X_1 and X_2 with $Ef_i(X_i) = 0$ for $i = 1, 2$. The challenge lies in finding suitable parametric representations for the smooth functions, and to control and choose the proper degree of smoothness.

Packages in R for fitting GAM models: **gam**, **mgcv** and **gss**. The **gss** package is a comprehensive implementation of the general smoothing spline approach to modeling described in the monographs by Wahba (1990) and Gu (2002). The underlying modeling approach depends on the notion of ANOVA decompositions of functions, which is different from the regular gam functions met so far. The main fitting methods for GAM are the packages gam and mgcv.

```
# Both functions are called 'gam'.
# Make sure you are using the desired one.
# Useful trick: always specify the library name before the function:
?gam::gam
?mgcv::gam
```

16.5.1 Fitting GAM with gam::gam

The model fitted by the function `gam::gam` is

$$g(E(Y|X)) = \beta_0 + \sum_j X_j\beta_j + \sum_j f_j(X_j).$$

The function `gam::gam` uses the backfitting algorithm (or Gauss-Seidel iteration) to combine different smoothing or fitting methods. The built-in methods are local linear/polynomial regression as discussed and smoothing splines as discussed in Section 3.1. The backfitting method fits the additive model by iteratively smoothing partial residuals. The algorithm separates the parametric from the nonparametric part of the fit, and fits the parametric part using weighted linear least squares within the backfitting algorithm.

- Usual arguments similar to glm(): model formula, family, dataset etc. Supporting functions such as residuals(), gam::predict.gam() and hatvalues() work for the fit object.

- For additive predictors, choose either s() for cubic smoothing spline, or lo() for local regression (loess) as the associated smoother, where you can specify degrees of freedom for smoothing spline, or degree of polynomial to be fit for local regression. For multi-dimensional smoothing, feed the corresponding predictors to lo(), e.g. lo(A,B) would represent a 2D smoothing regression surface.

- Check out the details by

  ```
  ?gam::s
  ?gam::lo
  ```

- Interactions do not work as in other packages Only parametric interactions are fitted, when models have both parametric and nonparametric parts. You need to define the corresponding covariate outside the function to do so.

- If predictors are not fed to the smoothing functions, the same linear effects as in GLM are fitted instead, resulting in a Generalized Additive Partial Linear Model (GAPLM).

- In the nonparametric ANOVA table, if some f_j is not significant, it means we can fit X_j with linear effect only (i.e., in the formula when implementing `gam::gam()`, put the name of X_j without any smoothing function such as `s()` and `lo()`).

- Parametric ANOVA tables in `gam::gam()` include the inference for β_0 and also the slopes $\beta_1, \ldots, \beta_p$.

Example

```
data(trees) # n=31, 2 predictors
gamgam1 <- gam::gam(Volume~s(Height,4)+s(Girth,5),
                    family=Gamma(link=log),data=trees)
# smoothing spline with df=4 for Height and df=5 for Girth.
summary(gamgam1)
plot(gamgam1) # plot the fitted smoothing function. See Figure 1.
```

```
Anova for Parametric Effects
             Df Sum Sq Mean Sq F value    Pr(>F)
s(Height, 4)  1 3.4475  3.4475  468.67 7.647e-16 ***
s(Girth, 5)   1 4.5814  4.5814  622.82 < 2.2e-16 ***
Residuals    21 0.1545  0.0074

Anova for Nonparametric Effects
             Npar Df  Npar F  Pr(F)
(Intercept)
s(Height, 4)       3 0.61277 0.6143
s(Girth, 5)        4 3.05267 0.0395 *
```

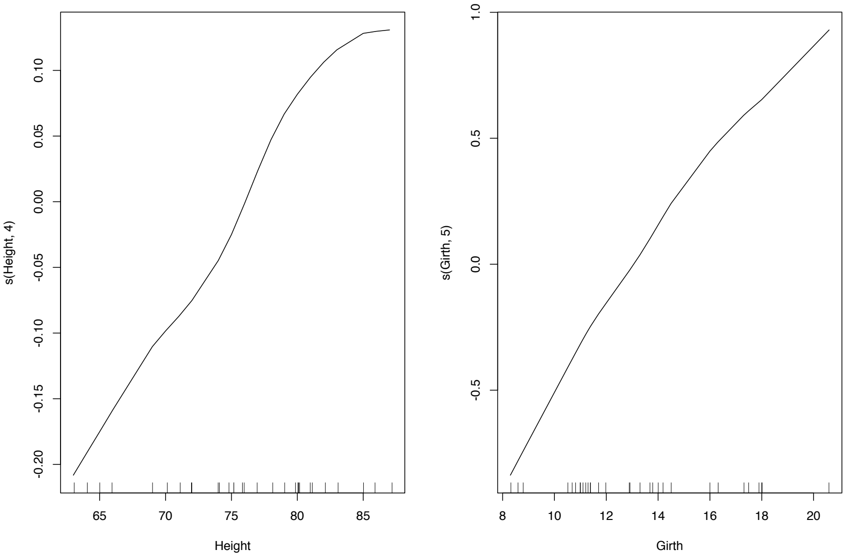


Figure 3: Fitted additive model *gamgam1* for trees data with smoothing splines using gam::gam.

```
gamgam2 <- gam::gam(Volume~lo(Height,degree=1)+s(Girth,5),
              family=Gamma(link=log),data=trees)
# local linear regression for Height,
# and smoothing spline with df=5 for Girth
```

```
summary(gamgam2)

plot(gamgam2) # See Figure 2.


Anova for Parametric Effects

                          Df Sum Sq Mean Sq F value    Pr(>F)

lo(Height, degree = 1)  1.00 3.4478  3.4478  465.99 6.861e-16 ***

s(Girth, 5)             1.00 4.5673  4.5673  617.29 < 2.2e-16 ***

Residuals              21.15 0.1565  0.0074


Anova for Nonparametric Effects

                       Npar Df Npar F   Pr(F)

(Intercept)

lo(Height, degree = 1)     2.9 0.5135 0.66849

s(Girth, 5)                4.0 3.2035 0.03333 *
```

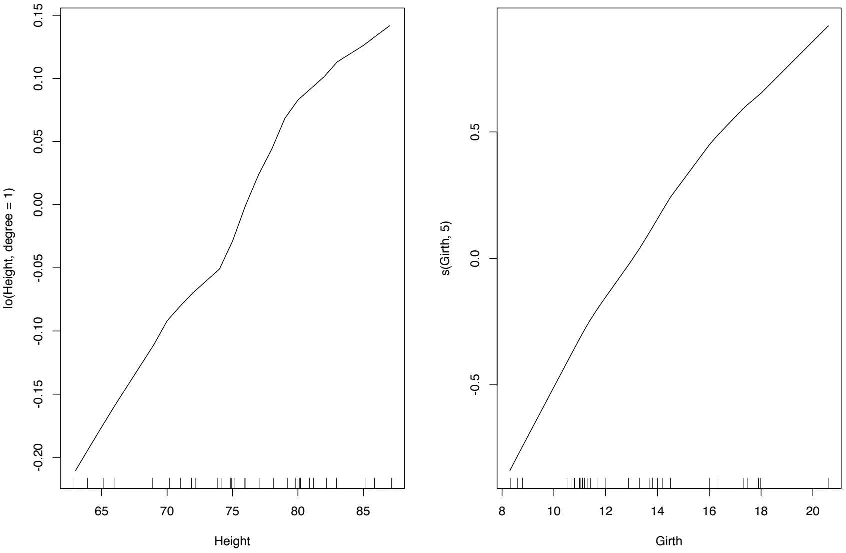


Figure 4: Fitted additive model *gamgam2* for trees data with local linear regression and smoothing splines using gam::gam.

```
gamgam3 <- gam::gam(Volume˜s(Height,4)+Girth,family=Gamma(link=log),data=trees)
```

```
# Fitting a GAPLM with linear effect on Girth
summary(gamgam3)
plot(gamgam3) # See Figure 3

Anova for Parametric Effects
             Df Sum Sq Mean Sq F value    Pr(>F)
s(Height, 4)  1 3.5047  3.5047  384.94 < 2.2e-16 ***
Girth         1 4.5868  4.5868  503.80 < 2.2e-16 ***
Residuals    25 0.2276  0.0091

Anova for Nonparametric Effects
             Npar Df Npar F Pr(F)
(Intercept)
s(Height, 4)       3 1.5112 0.236
Girth
```

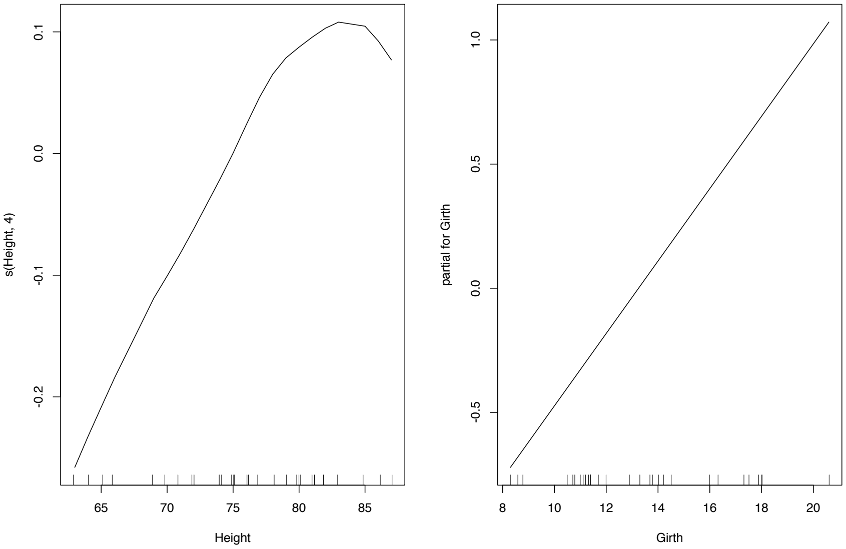


Figure 5: Fitted GAPLM *gamgam3* for trees data with smoothing splines for Height and linear effect for Girth using gam::gam.

16.5.2 Fitting GAM with mgcv::gam

The model fitted in `mgcv::gam()` is

$$g(E(Y|X)) = \beta_0 + \sum_j f_j(X_j).$$

Instead of using certain smoothers with specified degree of smoothness, **mgcv::gam** uses penalized regression splines (P splines) for smoothing, and fits the model together with smoothing parameters by GCV, REML (reduced maximum likelihood) and Bayesian methods.

- Similar arguments as in glm(). Supporting functions such as residuals(), mgcv::predict.gam() work for the fit object.

- In model formula, always use s() to denote the predictor(s) treated as additive effects. Here, s() does not mean using smoothing splines as in gam::gam function, but is just a indicator of the additive effect. Multi-dimensional smoothing is also supported by the function, e.g. s(A, B).

- Other useful options for s() include:

  - **bs**: a two letter character string indicating the (penalized) smoothing basis to use. (eg "tp" (default) for thin plate regression spline, "cr" for cubic regression spline).

  - **k**: the dimension of the basis used to represent the smooth term. The default (k=-1) depends on the number of variables that the smooth is a function of.

  Check out more details by **?mgcv::s**.

- The option **method** specifies the optimization method used. Default is "GCV.Cp", which performs GCV on Mallows' Cp, and we can also use "REML". For more choices, check with ?mgcv::gam.

- As mentioned previously, mgcv::gam manages to automatically determine the degree of smoothness by selecting smoothing parameters, based on Bayesian arguments, and this is very different from gam::gam, so the results are also different.
- Confidence/credible bands are available for visualization, based on Bayesian arguments.
- In nonparametric ANOVA tables, if some f_j is not significant, it means X_j should not be included in the model.
- Parametric ANOVA tables only have the inference for β_0 because there is no slope coefficient in this model.

Example

```
mgcvgam1 <- mgcv::gam(Volume~s(Height)+s(Girth),
          family=Gamma(link=log),data=trees)
summary(mgcvgam1)
plot(mgcvgam1, residuals=TRUE, shade = TRUE)
# confidence bands with observed scatter plot. See Figure 4.

# use the trees data for illustration as well
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.27570    0.01492   219.6   <2e-16 ***

Approximate significance of smooth terms:
            edf Ref.df      F  p-value
s(Height) 1.000  1.000  31.32 3.92e-06 ***
s(Girth)  2.422  3.044 219.28  < 2e-16 ***

mgcvgam2 <- mgcv::gam(Volume ~s(Height,bs="cr",k=15)+s(Girth, bs="cr", k=15),
          family=Gamma(link=log),data=trees)
```

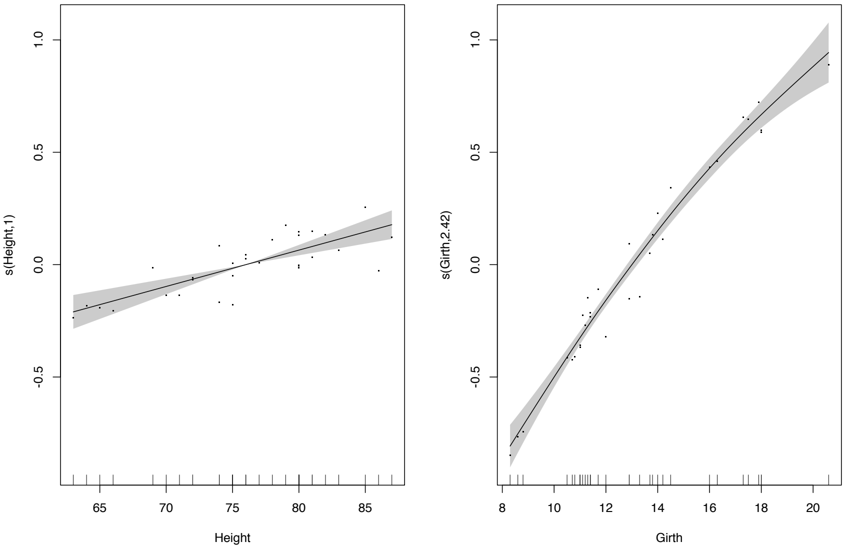


Figure 6: Fitted GAM *mgcvgam1* for trees data with default smoothing options in mgcv::gam.

```
# k specifies number of basis functions, and we use cubic splines here.
summary(mgcvgam2)
plot(mgcvgam2, residuals=TRUE, shade=TRUE) # See Figure 5.

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.273093   0.009713     337   <2e-16 ***

Approximate significance of smooth terms:
             edf Ref.df       F p-value
s(Height) 12.555  13.32   5.497  0.0101 *
s(Girth)   9.702  10.74 103.570  <2e-16 ***

mgcvgam3 <- mgcv::gam(Volume ~ s(Height)+Girth,
                      family=Gamma(link=log),data=trees) # Fit a GAPLM
summary(mgcvgam3)
```

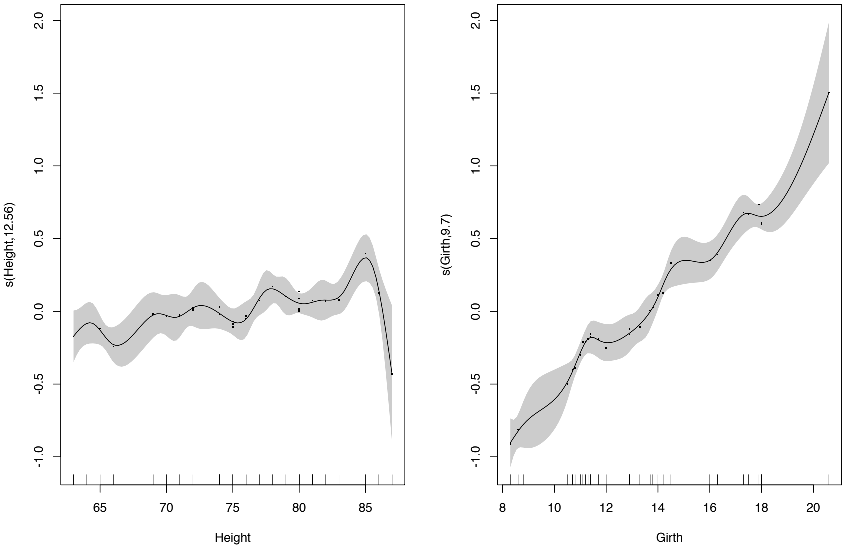


Figure 7: Fitted GAM *mgcvgam2* for trees data with specified number of cubic spline basis functions in mgcv::gam. (Is it a good fit? Why?)

```
plot(mgcvgam3, residuals=TRUE, shade=TRUE, all.terms = TRUE)

# See Figure 6.


Parametric coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.350148   0.087232   15.48 4.99e-15 ***

Girth       0.145412   0.006458   22.52  < 2e-16 ***


Approximate significance of smooth terms:

            edf Ref.df     F  p-value

s(Height) 1.737  2.166 13.11 6.11e-05 ***
```

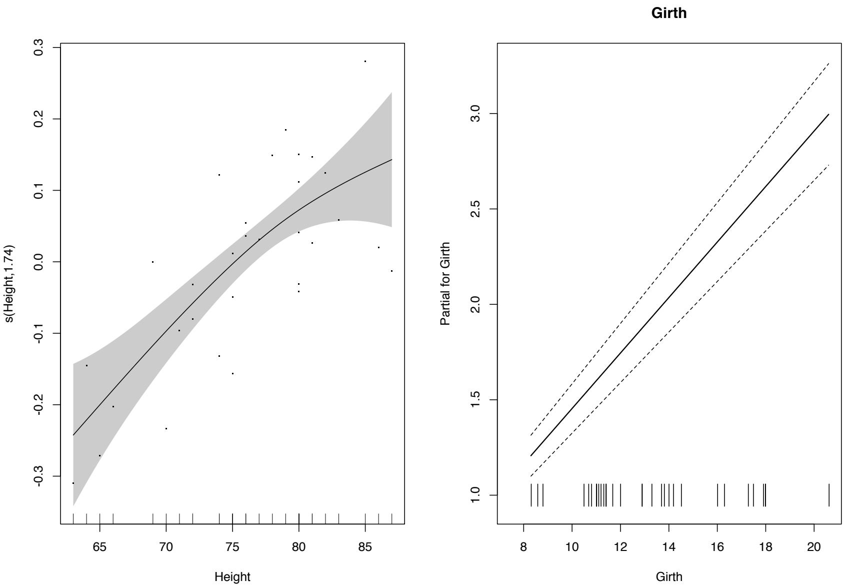


Figure 8: Fitted GAPLM *mgcvgam3* for trees data with default smoothing options for Height and linear effect for Girth via mgcv::gam.

17 Classification

17.1 Classification rules

Here we assume that the data are associated with two groups, G0 and G1. Examples occur in genetics where a gene expression pattern could be a signature for a cell cycle related gene or alternatively for a gene that is not cell cycle related. Or we would like to pinpoint gene expressions that are related to clinical diagnosis or prognosis. Observations consist of predictor vectors X (eg, vectors of gene expression) and responses Y are group membership indicator, with $Y = 0$ for group G0 and $Y = 1$ for group G1, and one typically has a *training data set* (X_i, Y_i), $i = 1, \dots, n$.

The task is to classify a subject with predictor vector X into G0 or G1. The Bayes rule of optimal classification is to classify according to

$$\text{argmax}_{k\in\{0,1\}}\, P(Y = k \mid X), \quad k = 0, 1. \tag{17.1}$$

Assuming marginal probabilities $\pi_1 = P(Y = 1)$, $\pi_0 = P(Y = 0)$, the prior probabilities of an arbitrary subject belonging to groups $G1$ or $G0$, $\pi_1 + \pi_0 = 1$, the Bayes formula yields

$$P(Y = k|X = x) = \frac{f_k(x)\pi_k}{f_0(x)\pi_0 + f_1(x)\pi_1}, \quad k = 0, 1, \tag{17.2}$$

where f_0, f_1 are the pdf (or pmf) of the covariates X for the subjects belonging to group G0 resp. to group G1.

The marginal probabilities π_0, π_1 are easy to estimate from the training set, using $\hat{p}_1 = \frac{n_1}{n}, \hat{p}_0 = \frac{n_0}{n}$, where n_0, n_1 are the numbers of subjects in the training set belonging to G0 resp. G1. Alternatively, they can be prespecified. If the densities f_0, f_1 can be reasonably well estimated, estimates can be substituted and this leads to the optimal classifier. Kernel density estimators can be employed if the data are continuous and low-dimensional (see Example (1) below).

The classification rule is: Classify into G1 if

$$\frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} = \frac{f_1(x)\pi_1}{f_0(x)\pi_0} > 1, \tag{17.3}$$

with estimates substituted for f_0, f_1, π_0, π_1, and otherwise classify into group G0. Sometimes work with the equivalent version: Classify into G1 if

$$\log \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} > 0. \tag{17.4}$$

Examples:

(1) *Use nonparametric density estimators to obtain f_0, f_1.* Given a sample $X_1, \ldots, X_n$ in $\mathcal{R}^1$, one can use kernel density estimates

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{17.5}$$

where K is a kernel function as discussed before (here usually chosen as a pdf with bounded or unbounded support) and h is a bandwidth (smoothing parameter), which controls the degree of smoothing and the trade-off between variance and bias. Extensions to multivariate cases are straightforward, but *boundary corrections* are a major issue (also for the univariate case).

An alternative implementation is the smoothing of histograms with local least squares (implemented e.g. in HADES – hazard and density estimation, in matlab at
http://anson.ucdavis.edu/ mueller/data/hades.html)
This approach works if $X \in \mathcal{R}^p$ for $p = 1, 2$, X is continuous and the sample size n is not small. For larger dimensions p this method and indeed any fully nonparametric method for density estimation will run into the *curse of dimensionality.*

Estimators $\hat{f}$ in (17.5) depend on a bandwidth h, so this and also many other classifiers require a tuning parameter, which is best chosen by CV from a *tuning set*, a subset of the training set. Then the chosen tuning parameter is used for the remainder of the training set to determine the optimal classifier, which then in turn is evaluated on the test set.

Note: In a simulation setting, the test set should be selected very large, which is always possible as it is simulated. This stands in contrast to data analysis situations, where often there is no clearly defined test set and one has to split the sample artificially to create training and test sets.

(2) *Gaussian assumptions.* Here one assumes that f_0, f_1 are multivariate Gaussian, i.e.,

$$f_k(x) = (2\pi)^{-p/2}\det(\Sigma_k)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)\right).$$

If $\Sigma_k = \Sigma,\ k = 0, 1$, then the classification rule becomes

$$\log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} = \log \frac{\pi_1}{\pi_0} + (x - \frac{1}{2}(\mu_1+\mu_0))^\top \Sigma^{-1}(\mu_1-\mu_0)), \tag{17.6}$$

which is a linear function in x, the value of which is compared to 0 to carry out the classification. Applying (17.4) leads to *Linear Discriminant Analysis (LDA).*

A second case of interest occurs when $\Sigma_0 \neq \Sigma_1$. In this case, the classification rule is based on

$$\begin{aligned}\log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} &= \log \frac{\pi_1}{\pi_0} - \frac{1}{2}\log \frac{\det(\Sigma_1)}{\det(\Sigma_2)} \\ &\quad -\frac{1}{2}[(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)].\end{aligned}$$

One then applies the rule (17.4). This is a function that is quadratic in x, therefore this method is called *Quadratic Discriminant Analysis (QDA).*

When implementing linear and quadratic discrimination rules, means, covariances and prior probabilities need to be estimated from the training set. If the predictor dimension p is relatively large, the number of these parameters is very large, as it is quadratic in p. This causes problems of stability and variance of the estimates. Sparse selectors and shrinkage estimates of the covariance matrices are then advantageous.

17.2 Binomial discriminant analysis

Observe that optimal classification can be based on the classification function

$$\begin{aligned}\log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} &= \log \frac{P(Y=1 \mid X=x)}{1-P(Y=1 \mid X=x)} \\ &= \operatorname{logit}[P(Y=1 \mid X=x)] = \operatorname{logit}[E(Y \mid X=x)]. \quad (17.7)\end{aligned}$$

This suggests *logistic discriminant analysis*, setting

$$\operatorname{logit}[E(Y \mid X=x)] = x^\top \beta = \eta. \quad (17.8)$$

This is implemented by estimating β from the training set and using a logistic regression for the binary regression data $(X_i, Y_i)_{i=1,\ldots,n}$.

More generally, one can use a GLM and set $g(E(Y \mid X = x) = \eta$ for a link function g that is suitable for a binomial regression, i.e., the inverse of a cdf. In either case, the classification rule is to assign to G1 if $\text{logit}[E(Y \mid X = x)] = \eta > 0$ or if, observing that $\text{expit}(0) = 1/2$,

$$\mu = P(Y = 1 \mid X) > \frac{1}{2}.$$

Logistic discriminant analysis makes fewer assumptions than linear discriminant analysis (both have linear discriminant functions but these are differently estimated) and is often preferred.

The quality of the classification depends on the choice of the link function and the validity of the basic model assumptions. One can also run logistic discriminant analysis with a quadratic predictor, by replacing the linear predictor $\eta = \sum_{j=1}^{p} \beta_j x_j$ with

$$\tilde{\eta} = \sum_{j=1}^{p} \beta_j x_j + \sum_{j,k=1, j \le k}^{p} \gamma_{jk} x_j x_k.$$

Nonparametric Variants of Binomial Classification

(a) GAM:

For continuous predictors $X_1, \dots, X_p$, fit $\text{logit}(E(Y|X = x)) = \alpha + \sum_{j=1}^{p-1} g_j(x_j)$, with component functions $g_1, \dots, g_{p-1}$, then classify to G1 when $\mu = P(Y = 1 \mid X) > \frac{1}{2}$.

(b) Smoothing:

For $p \le 3$ (low-dimensional predictor case), one can use smoothing methods to obtain estimates of $h(x) = E(Y \mid X = x) = P(Y = 1 \mid X = x)$ from the scatterplot (X_i, Y_i) and then obtain the estimate $\text{logit}(\hat{h}(x))$ for $\text{logit}(P(Y = 1 \mid X = x))$.

17.3 Assessment of classifiers

In order to assess the quality of a classifier, one usually divides the available data into a *training set* and a *test set*. The training set is used to fit the model and, if applicable, a subset, the *tuning set*, is used to choose tuning parameters. The training of the model in the training set aims to minimize the *misclassification rates*

$$p_1 = P(\text{ classified into G0} \mid \text{G1})$$

and

$$p_0 = P(\text{ classified into G1} \mid \text{G0}),$$

and in particular the overall misclassification rate $p_m = p_1\pi_1 + p_0\pi_0$. Misclassification rates are estimated by the corresponding relative frequencies. These estimates are overly optimistic and biased since the model is specifically adapted to the training data. For new data, it will perform worse.

First optimize the model by minimizing the apparent error rate for the training data. Auxiliary parameter selection (smoothing or other tuning parameters and also predictor selection) preferrably is done on a separate tuning set. The optimized model is then applied to the test data. The predictors for the test data are used to obtain the classification, and relative frequency estimates are computed for the misclassification rate on the test data.

Problem: Often no test data available, only training set. Options: Can split available data into training and test data set (artificially and randomly). Then repeat random splits many times and average over misclassification rates: k-fold cross-validation (CV) removes a fraction of $1/k$ of the data each time as test set and uses the rest as training set, to be repeated k times, such that the test sets are disjoint. Or use *one-leave-out cross-validation* where $k = 1$ and each observation is left out in turn, which produces the CV misclassification rate:

$$CVM = \frac{1}{n}\sum_{i=1}^{n} |\hat{Y}^{(-i)} - Y_i|$$

where $\hat{Y}^{-i}$ is the estimated group membership indicator obtained from the predictors X_i plugged into the model that has been fitted omitting the data for the i-th subject (X_i, Y_i).

Usually will report results of classification in a 2×2 table ("confusion matrix")

	Classified	
True	G0	G1
G0	n_{00}	n_{01}
G1	n_{10}	n_{11}

The table will be reported for test set counts or averages of cross-validation test set counts in which test and training sets do not overlap.

Misclassification rates (apparent or cross-validated):
Overall misclassification rate: $(n_{01} + n_{10})/(n_{01} + n_{10} + n_{00} + n_{11})$
Misclassification rate for G1: $n_{10}/(n_{10} + n_{11})$
Misclassification rate for G0: $n_{01}/(n_{00} + n_{01})$.

17.4 LAB 7: Classification and Prediction Error

17.4.1 Discriminant Analysis

There are many classification methods implemented in R. We will introduce functions to fit Linear Discriminant Analysis and Quadratic Discriminant Analysis discussed in the lecture notes. It is important to keep in mind that both methods rely on assuming the joint distribution of predictors under consideration is *Gaussian*. Therefore, these approaches do not work for categorical predictors. When these are present *logistic discrimination* is the method of choice.

Linear Discriminant Analysis Use the function

```
library(MASS)
?lda
```

to perform linear discriminant analysis (Fisher's linear discriminant).

- A grouping factor/class factor is required in the model formula.
- "CV": a logical option indicating whether leave-one-out cross validation is performed, default: "FALSE".
- "prior": can be used to specify prior probabilities of class membership, default: taken as sample proportion.
- "subset": a index vector specifying the cases to be included in the training set. If not specified, the whole data will be treated as training.

Example

```
Iris <- data.frame(rbind(iris3[,,1], iris3[,,2], iris3[,,3]),
                   Sp = rep(c("s","c","v"), rep(50,3)))
# n=150, p=4, N_s = N_c = N_v = 50
train <- sample(1:150, 75)

z <- lda(Sp ~ ., Iris, prior = c(1,1,1)/3, subset = train)
pred <- predict(z, Iris[-train, ])$class
test <- Iris$Sp[-train]
table(pred, test)
# To get the misclassification rate
misrate <- length(which(test != pred))/(150-length(train))

    test
pred  c  s  v
   c 20  0  2
   s  0 27  0
   v  1  0 25
```

```
zcv <- lda(Sp ~ ., Iris, prior = c(1,1,1)/3, CV = TRUE)
# it contains CV-predicted class label and fitted probabilities
pred_cv <- zcv$class
table(pred_cv, Iris$Sp)

pred_cv  c  s  v
      c 48  0  1
      s  0 50  0
      v  2  0 49
```

Quadratic Discriminant Analysis Under the same package, we can use the function

```
?qda
```

to perform quadratic discriminant analysis. Similar options are available as in the "lda" function.

Example

```
tr <- sample(1:50, 25)
train <- rbind(iris3[tr,,1], iris3[tr,,2], iris3[tr,,3])
test <- rbind(iris3[-tr,,1], iris3[-tr,,2], iris3[-tr,,3])
cl <- factor(c(rep("s",25), rep("c",25), rep("v",25)))
# class labels are the same for training and testing
zq <- qda(train, cl)
predq <- predict(zq,test)$class
table(cl, predq)
# To get misclassification rate
misrateq <- length(which(cl != predq))/nrow(test)



   predq
cl   c  s  v
```

```
  c 24  0  1
  s  0 25  0
  v  1  0 24
```

```
zqcv <- qda(train, cl, CV = TRUE)
pred_qcv <- zqcv$class
table(cl, pred_qcv)
```

```
   pred_qcv
cl   c  s  v
  c 24  0  1
  s  0 25  0
  v  0  0 25
```

17.4.2 Cross Validation for Model Selection and Model Evaluation

In applications involving classification or prediction problems, we often evaluate or select the final model by comparing the misclassification rate or MSE etc. It is essential to obtain fair and objective assessments for the models under consideration. Cross validation is a popular approach to help us out in these situations. The basic idea is to split data to avoid using the same data twice for both model fitting and assessing (a.k.a double-dipping), which often results in overly optimistic conclusions or overfitting.

Cross validation can be used in two stages of data analysis:

- model selection/tuning for each individual model (you can turn to other methods such as AIC/BIC, LRT etc. for model selection as well);
- model evaluation/assessment for all candidates of final model.

In general, to perform rigorous cross-validation, the dataset is divided into three parts:

- training set, where model is established and developed.

- tuning set, where tuning parameters for the model are selected via optimization (Model Selection); for example smoothing parameter in nonparametric based discriminant analysis or predictor selection.
- test set, where we assess the performance of the tuned model (Model Assessment).

For tuning parameter estimation/selection, it is common to pool together the training and tuning and do a K-fold cross validation (think about why tuning procedure should not be done with the test set). However, for some models, e.g., GLM, a tuning parameter may not be present, and cross-validation for model selection/assessment is then done over the whole dataset.

Instead of implementing CV procedures explicitly, we can utilize the built-in CV functions in R. For GLM, there are two sets of R commands available to perform cross validation automatically.

Elastic Net GLM for Model Selection An elastic net GLM takes a penalized maximum likelihood approach, with a penalty term involving a mixture of l_1 norm and l_2 norm of the model coefficients (β_j),

$$l_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|, \quad l_2(\boldsymbol{\beta})^2 = \sum_{j=1}^{p} \beta_j^2, \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T.$$

This allows the model to select useful predictors when p is relatively large and manual model selection is not feasible. Tuning parameter is the regularization parameter λ. We will use the package "glmnet".

```
library(glmnet)
?glmnet # fit a model with a set of lambda
```

- uses the design matrix "x" and the response vector "y" instead of a model formula.
- need to specify a exponential family distribution using the "family" option as in "glm()".

- "alpha": the weight between l_1 and l_2 norm, yielding the penalty as

  $$p_\alpha(\boldsymbol{\beta}) = \alpha l_1(\boldsymbol{\beta}) + \frac{1-\alpha}{2} l_2(\boldsymbol{\beta})^2, \quad \alpha \in [0,1].$$

  Here, $\alpha = 1$ is the default choice, equivalent to the Lasso set-up, which features the ability to shrink the coefficients of unfavorable predictors to 0. When $\alpha = 0$, it corresponds to the ridge type of penalty.

- "lambda": the multiplier in front of the penalty, and the optimal coefficients are selected as

  $$\boldsymbol{\beta}^* = \text{argmin}_{\boldsymbol{\beta}} \left(-l(\boldsymbol{\beta}, y) + \lambda p_\alpha(\boldsymbol{\beta}) \right).$$

  By default, we do not need to specify this value, which results in a default selection of 100 candidates for λ, and all models are fitted in one run (the number of regularization parameters considered can be specified using the "nlambda" option). Alternatively, the regularization parameter can be specified by the user.

- Only model fitting is performed using this function, no assessment criterion is calculated. For more details, please refer to the documentation.

```
?cv.glmnet # K-fold cross validation for a saturated model
```

- input data and exponential family arguments are the same as glmnet().

- "type.measure": the type of loss function to be used for the cross-validation optimization criterion. It will be calculated for all models corresponding to different λ, and stored in the model object. There are currently 5 choices, not all available for all models. So we need to specify the proper one for our data:

  - "deviance": deviance for logistic and poisson regression;
  - "class": misclassification error, applies to binomial and multinomial regression only;
  - "auc": for two-class logistic regression only, and gives area under the ROC curve;

  - "mse" (mean squared error) or "mae" (mean absolute error): deviation from the fitted mean to the response, applies to all models except "Cox" model in survival analysis.

- the output object contains
  - "lambda": vector of regularization parameters that are considered; "cvm": corresponding cross-validated errors for these regularization parameters.
  - "lambda.min": the value of λ attaining the smallest loss in terms of cross-validation;
  - "glmnet.fit": the model fitting results for all models corresponding to all λ considered. See the example below.

Example

```
set.seed(1010)
n=1000;p=100
nzc=trunc(p/10) # nonzero coefficient columns
x=matrix(rnorm(n*p),n,p) # design matrix
beta=rnorm(nzc) # nonzero coefficients
fx= x[,seq(nzc)] %*% beta # linear predictor
px=exp(fx)
px=px/(1+px) # true probability
ly=rbinom(n=length(px),prob=px,size=1) # binary response vector
set.seed(1011)
cvob_bin=cv.glmnet(x,ly,family="binomial") # default: alpha=1 (lasso)
plot(cvob_bin)
title("Binomial Family",line=2.5) # See Figure 1.
# select the optimal regularization parameter lambda
opt_lambda_idx = which(cvob_bin$lambda == cvob_bin$lambda.min)
opt_lambda = cvob_bin$lambda[opt_lambda_idx]
### check first 15 of the corresponding coefficients
head(cvob_bin$glmnet.fit$beta[,opt_lambda_idx], n = 15)
```

```
         V1          V2          V3          V4          V5          V6
-0.22345445  0.74778070 -0.38579899 -0.27145031 -0.09248802  0.62204227
         V7          V8          V9         V10         V11         V12
 0.59975084 -1.38327355  1.22586261  0.13767835  0.00000000  0.00000000
        V13         V14         V15
 0.00000000  0.01825499 -0.06855708
# We see that some coefficients are shrunk to zero.

# check the corresponding number of nonzero coefficients
which(cvob_bin$glmnet.fit$beta[,opt_lambda_idx] > 0)

  V2   V6   V7   V9  V10  V14  V17  V19  V25  V27  V28  V29  V35  V38
   2    6    7    9   10   14   17   19   25   27   28   29   35   38
 V44  V49  V59  V73  V77  V97 V100
  44   49   59   73   77   97  100

# Fit with optimal lambda:
# DO NOT do this in practice, as it is double-dipping.
# Fit the model on TEST SET instead.
doubledippingfit = glmnet(x,ly,family="binomial",lambda=opt_lambda)
```

Embedded Cross-Validation in R For many popular methods including linear regression and GLM, R provides a convenient way to do cross validation.

```
library(cvTools)
library(boot)
?cv.glm
```

- “glmfit”: a glm model object fitted to the whole dataset.

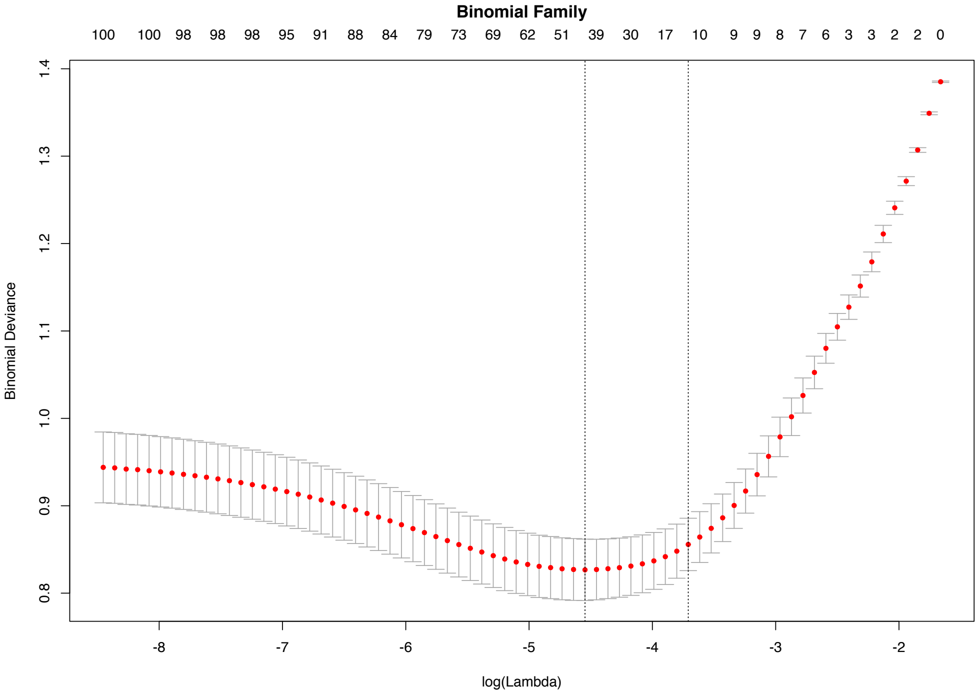


Figure 9: Plot of cross validation errors against candidate regularization parameters for an elastic net logistic regression example. The numbers at top are the numbers of predictors left as lambda increases.

- “data”: the data used for establishing the “glmfit”. Can be a data frame or a matrix.
- “K”: folds of the cross validation, with default: leave-one-out cross validation. Observations are randomly allocated to each fold.
- “cost”: the default is the average squared error function. Users can specify their own cost function, which must be a function of two vector arguments. The first argument corresponds to the observed responses and the second argument to the predicted or fitted responses from the glm. The function must return a non-negative scalar value.
- In the returned object, the component “delta” includes the raw and adjusted cross validated error. The adjustment is designed to compensate for the bias introduced by not using leave-one-out cross-validation.

Example

```
data(mammals, package="MASS")
```

```
mammals.glm <- glm(log(brain) ~ log(body), data = mammals)

cv.err <- cv.glm(mammals, mammals.glm)$delta # leave-one-out CV

cv.err

[1] 0.4919 0.4917

cv.err.6 <- cv.glm(mammals, mammals.glm, K = 6)$delta # 6-fold CV

cv.err.6

[1] 0.4885 0.4864
```

17.5 PROBLEM SET 5

1. For a Gamma regression model, discuss the advantages and disadvantages of the link function choices $g(\mu) = \log(\mu)$ and $g(\mu) = 1/\mu$.

2. Consider the *multiplicative error model* $y = \mu(\epsilon+1),\ \ \epsilon > -1,\ \ \mu > 0,\ \ E\epsilon = 0,\ \ \mathrm{var}(\epsilon) = \sigma^2$. (a) Show that this is a constant coefficient of variation model. (b) If you know that data have been generated in this way, what are your options for implementing this model? Discuss at least two options and their pros and cons.

3. Derive the quasi-likelihood for a constant-coefficient-of-variation model and link function g.

4. Provide three models which can be characterized as constant-coefficient of variation models. For each model give its name, and the relation of the responses y to the predictors as well as distribution of y if specified in the model.

5. Show that the log transformation approximately stabilizes the variance but leads to an off-set in the mean for constant coefficient of variation models.

6. Derive the quasi-information matrix i_β by direct evaluation of $E(\frac{\partial U}{\partial \beta})$ and show that quasi-information equals information in the case of an underlying exponential family, i.e., $D^T V^{-1} D/\sigma^2 = X^T W X$.

7. For a quasi-normal model with identity link, obtain (a) quasi-likelihood and quasi-score as a function of the parameter vector β, and (b) the quasi-information. How can the quasi-information be employed in model fitting?

8. In a nutritional dietary study, in which n_1 women from Japan and n_2 women from the US are enrolled, an investigator is interested to demonstrate that the effect of dietary daily saturated fat intake (FAT) on breast cancer is the same in the U.S. and in Japan. FAT is measured as a daily average food intake in terms of fat calories. Assume the observation Y for each woman is whether breast cancer occurs $Y = 1$ or does not occur

$Y = 0$ in a fixed time period. An additional covariate is age, which is assumed to have the same effect for US and Japanese women. We assume that Y is observed for all participants without censoring or otherwise missing values.

(a) Set up a suitable GLM and specify all components. (Note: The model needs to allow for the presence of a difference in the effect of the amount of daily fat intake between US and Japanese women).

(b) Discuss the choice of the link function. Assume that it is not clear one should use a logit link and there may be some evidence for an asymmetric link function. How could you proceed?

(c) Formulate the null hypothesis and alternative you would set up to find out whether the investigator's hypothesis is correct. Would you plan to discuss the set-up of the hypotheses with the investigator? How would these discussions be reflected in the formulation of the null hypothesis?

(d) Sketch in a graph how you would carry out the test (provide a rough graph that shows the region of the null hypothesis and a suitable confidence region). Label all elements that appear in your graph and indicate how you can obtain a p-value. Detailed formulas are not required.

(e) Assume in a second analysis only the US women are studied. The investigator is interested to find out whether the effect of age and fat intake is linear or not. How could you study this using an embedding technique? Assume that a formal test is desired by the investigator. How can you proceed?

(f) Following up on (e), the investigator wants to you consider that the relation of the response Y with the predictors AGE and FAT could be complex. How can you study the relation with nonparametric regression? Specify the model as a conditional probability and provide an explicit formula for an estimation method. Specify how you select the tuning parameter.

9. Derive a quasi-Gamma model with a log link. First, write the variance function. Then obtain, as a function of β, the quasi-score, the quasi-likelihood and the covariance of the score.

10. Consider the model $E(Y|X) = \alpha + \sum_j g_j(X_j)$, with $E(g_j(X_j)) = 0$, where the predictors X_j are independent random variables. In this situation, calculate $E(Y|X_j)$ and use this result to find a simple and consistent fitting method that does not require backfitting. Discuss in which situations you might have independent predictors.

11. (a) Give a precise derivation of the Bayes formula and (17.2).

    (b) Derive (17.4).

    (c) Derive (17.6).

(d) Derive (17.7).

(e) Is the logistic classifier a Bayes classifier?

(f) What is the effect of the marginal probabilities π_0, π_1 on the logistic classifier?

12. For the PIMA Indians diabetic study, for the following problems assume only the predictors labeled 2-6 and 8 (Plasma glucose concentration at 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), and age).

    (a) For the data with the reduced number of predictors, obtain a GLM with canonical link, selecting appropriate predictors and carrying out the diagnostics.

    (b) Fit a binary GAM regression model with canonical link. Display and interpret the selected additive functions f_j. Compare the GLM and the GAM fit for these data.

    (c) Would you replace some of the components of the GAM by linear functions? If so, fit linear functions for these components and smooth functions for the remaining components in a GPLAM and interpret the results.

13. For the data in the previous problem, compare the classification performance of the following four classifiers by 10-fold cross-validation: (a) logistic classifier with a linear predictor (b) logistic classifier with a quadratic predictor (c) Fisher's linear discriminant analysis (d) Quadratic discriminant analysis based on the normality assumption.

14. In a study on diabetes, one wants to investigate the relationship of patient characteristics with disease outcomes. Specifically, data on (a) glucose tolerance (GLUC, a level of glucose in the blood and thus a continuous variable) and (b) Age (AGE) are obtained for a sample of 200 randomly selected patients. The outcome variable is whether these patients have diabetes (yes or no).

    (a) Write all components for analyzing these data for a GLM and a GAM, including constraints for the GAM.

    (b) Explain why GAM overcomes the "curse of dimensionality", which affects other nonparametric approaches, even if the number of predictors is large.

    (c) When you present and interpret the results from GAM, which items do you include and how are they interpreted?

    (d) Assume that gender (GEN) is included as an additional predictor. Describe a model that can handle this additional predictor while retaining general additive effects for the continuous predictors.

(e) When fitting a GLM to these data, how can you test whether there is an interaction between GEN and each of the other two predictors? Write the null hypothesis, the test statistic in the form of a general linear test and describe its distribution under H_0.

15. For the epileptic seizure data in chemo.dat, redo the original analysis by comparatively fitting the following four models: (1) Poisson model with overdispersion, as you did before; (2) a negative binomial model; (3) zero-inflated Poisson model; (4) zero-inflated negative binomial model.

    (a) Write these models and show the relevant results for these four models. Discuss the comparisons.

    (b) Is there evidence for zero-inflation?

    (c) Which among these models would you choose for your final analysis? Give detailed reasoning for your model choice.

    (d) Interpret your results for the final model chosen.

16. In the epilepsy data chemo.dat we have four consecutive two-week periods during each of which we count the number of epileptic episodes. Assume you want to predict the count of episodes in the fourth period from the three episode counts of the first three periods. Please formulate a model. Write null hypothesis, provide the test statistics and describe how you would carry out the test if you have access to an implementation of IWLS program. How would you devise a test to determine whether the influence of all three predictors on the outcome is the same.

17. Assume that in a randomized clinical trial patients with epilepsy are given either drug A or drug B. Here B is a new drug and the goal is to determine whether it is better in reducing the number of seizures compared to an established drug A. We count the number of epileptic seizures in a six month period while they are taking the drug. Gender and time since onset of the disease are also recorded.

    (a) Provide all components of a suitable GLM, where it is decided to use a Poisson regression with the canonical link. Write the main null hypothesis and alternative of interest to address the main question of the study.

    (b) Write the information matrix on which the approximate inference for the model is based; it suffices to specify a general element of the matrix. How is approximate inference obtained from this matrix?

    (c) How do we estimate over-dispersion? Provide details how it can arise by writing a model that leads to overdispersion. In the presence of overdispersion, is there an alternative to Poisson regression?

(d) Assume that it is found that some of the patients were misdiagnosed and do not actually have epilepsy. Accordingly zero seizures were recorded for these, but we do not know which patients these are. How can you address this situation? Provide details and all components of a suitably modified model. How can we determine the fraction of patients that do not have epilepsy?

18 Dose-Response Models

18.1 The threshold model

Applying a certain dose of a substance, a ”quantal” (0-1) response is observed for a subject.

Example: Want to establish dose-response relation between cigarette smoke exposure and lung cancer. Apply doses d of cigarette smoke to mice → observe whether tumor develops, yes/no.

Carcinogenesis, Multiple stage model:

	□	0	Healthy
Repair	↑ ↓	Damage:	Mutations, Gene deletions
	□	1	Precancerous
	↓	Damage	
	□	2	Cancerous
		Cancer stage	

A certain dose of a cancerogenic substance can be tolerated due to repair mechanisms: Each subject has individual (random) threshold X_i:

Threshold Model

If $d < X_i$: Damage can be repaired, no tumor
If $d > X_i$: Repair mechanisms are overwhelmed and a tumor develops

Paracelsus (1493-1541): *Sola dosis facit venenum – Only the dose makes poison.*

Dose-response relationships may help to establish causality; establishing "safe" doses in toxicology, "effective" doses in vaccine development.

Next assume thresholds $X_i \sim F$ for a cdf F. Then probability of a response (tumor develops) at dose d is

$$p(d) = P(X_i \leq d) = F(d).$$

We now assume there is a standardized version of F, which we denote by F_0:
If $EX = \mu, \operatorname{var}(X) = \sigma^2 > 0$, set $F_0(x) = P(\frac{X-\mu}{\sigma} \leq x)$. Then

$$\begin{aligned} F(d) &= P(X \leq d) = P(\frac{X-\mu}{\sigma} \leq \frac{d-\mu}{\sigma}) \\ &= F_0(\frac{d-\mu}{\sigma}) = F_0(\beta_0 + \beta_1 d) = F_0(\eta) \end{aligned}$$

if we let

$$\beta_0 = -\mu/\sigma, \qquad \beta_1 = 1/\sigma$$

and the linear predictor $\eta = \beta_0 + \beta_1 d$. This establishes a connection to binomial regression with link function

$$g \equiv F_0^{-1}.$$

In practice, it turns out that often

$$\log X_i \sim_{iid} F$$

provides a better modeling assumption. Then the linear predictor is chosen as $\eta = \beta_0 + \beta_1 \log d$.

18.2 Examples

(a) The Probit Model

$$\log X \sim N(\mu, \sigma^2)$$

$$p(d) = \Phi(\frac{\log d - \mu}{\sigma}) = \Phi(\beta_0 + \beta_1 \log d),$$

here Φ is the cdf of the standard normal distribution.

(b) The Logit Model

The logistic distribution with parameters (μ, τ) has pdf

$$f(x) = e^{(x-\mu)/\tau}/\{\tau(1+e^{(x-\mu)/\tau})^2\},$$

$X \sim f$ has $EX = \mu$, $\text{var}(X) = \pi^2\tau^2/3$, $\sigma = \tau\pi/\sqrt{3}$,
with $\log X \sim f$, $\eta = \beta_0 + \beta_1 \log d$, $p(d) = \exp(\eta)/(1+\exp(\eta))$,
where $\beta_0 = -\mu/\sigma$, $\beta_1 = 1/\sigma$, ignoring constants.

(c) The complementary log model

The Gumbel extreme value distribution with parameters (α, κ) has pdf

$$f(x) = \frac{1}{\kappa}\exp(\frac{x-\alpha}{\kappa})\exp[-\exp(\frac{x-\alpha}{\kappa})], \quad -\infty < x < \infty$$

with $X \sim f$, $\mu = \alpha + \kappa\gamma$, $\sigma = \kappa\pi/\sqrt{6}$ for $\gamma \approx 0.577$ (Euler's constant).

Then, ignoring constants,

$$\log(-\log(1-p(d))) = \beta_0 + \beta_1 d = \eta,$$

with $\beta_0 = -\alpha/\kappa$, $\beta_1 = 1/\kappa$, so the link function is

$$g(\mu) = \log(-\log(1-\mu)).$$

18.3 Dilution assays

Of interest is concentration of infective agents (bacteria, viruses etc.) in blood. This concentration cannot be directly measured; can only observe growth of samples, which are either at original or at diluted concentrations.

In a Petri dish (growth medium): observe growth yes/no. We assume the blood concentration is N agents/volume unit.

The 1:1 dilution assay:
The i-th assay has $\frac{N}{2^i}$ agents/volume. Assume the number of growth centers observed on a Petri plate inoculated with the i-th assay is Poisson distributed:

$$N_i \sim P(\alpha_i), \quad \alpha_i = \frac{N}{2^i}.$$

Then

$P(\text{growth observed for } i\text{−th dilution assay}) = 1 - P(\text{no growth observed}) = 1 - e^{-\alpha_i} = 1 - e^{-N/2^i}$.

Repeat the i-th assay m times, and obtain $\hat{p}_i = y_i/m$, where y_i is the number of plates where growth was observed among m trials. Then

$$\mu_i = E\hat{p}_i = 1 - e^{-\alpha_i}, \;\; \alpha_i = -\log(1-\mu_i),$$

$$\log(-\log(1-\mu_i)) = \log \alpha_i = \log N - i \log 2$$

Binomial regression model with complementary log-link:

$$\log(-\log(1-\mu_i)) = \beta_0 + \beta_1 i,$$

where $\beta_1 = -\log 2$ is known, β_0 unknown.
After fitting the model, $\hat{N} = e^{\hat{\beta}_0}$.

18.4 Estimating effective doses

$ED\alpha/LD\alpha$: effective/lethal doses such that $p(ED\alpha) = \alpha$, $0 < \alpha < 100$, i.e., $\alpha\%$ of the population will show a response at this dose. This means that $\alpha\%$ of the individual thresholds are smaller than this dose.

Example: Logit model:

$$\text{logit}(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 d; \quad \text{logit}(0.5) = 0,$$

$$\text{therefore} \quad ED50 = -\beta_0/\beta_1, \quad E\hat{D}50 = -\hat{\beta}_0/\hat{\beta}_1;$$

In general,

$$ED\alpha = (\log \frac{\tilde{\alpha}}{1-\tilde{\alpha}} - \beta_0)/\beta_1,$$

where $\tilde{\alpha} = \alpha/100$.

With general link function $g \equiv F^{-1}$, find d such that

$$F(\beta_0 + \beta_1 d) = \tilde{\alpha} \qquad d = (F^{-1}(\tilde{\alpha}) - \beta_0)/\beta_1,$$

or

$$ED\alpha = (F^{-1}(\tilde{\alpha}) - \beta_0)/\beta_1$$

$$\widehat{ED}\alpha = (F^{-1}(\tilde{\alpha}) - \hat{\beta}_0)/\hat{\beta}_1.$$

For log doses,

$$ED\alpha = \exp\{(F^{-1}(\tilde{\alpha}) - \beta_0)/\beta_1\}, \quad \widehat{ED}\alpha = \exp\{(F^{-1}(\tilde{\alpha}) - \hat{\beta}_0)/\hat{\beta}_1\}.$$

18.5 Asymptotic inference for ED_α

Set $\beta = (\beta_0, \beta_1) \in \mathcal{R}^2$. There exists a function h such that $E\hat{D}\alpha = h(\hat{\beta}_0, \hat{\beta}_1) = h(\hat{\beta})$ and

$$\sqrt{n}(\hat{\beta} - \beta) \to_D N_2(0, I^{-1}) \quad (n \to \infty).$$

From the GLM asymptotics, the relevant information matrix is $I = \lim_{n\to\infty} \frac{1}{n} X^T W X$.
By Taylor expansion, aiming at the δ-method,

$$\widehat{ED}\alpha = ED\alpha + (\hat{\beta}_0 - \beta_0)\frac{\partial}{\partial \beta_0} h(\beta) + (\hat{\beta}_1 - \beta_1)\frac{\partial}{\partial \beta_1} h(\beta) + o_p(|\hat{\beta}_0 - \beta_0| + |\hat{\beta}_1 - \beta_1|),$$

so that

$$\sqrt{n}(\widehat{ED}\alpha - ED\alpha) = \sqrt{n}(\hat{\beta}_0 - \beta_0)\frac{\partial}{\partial \beta_0}h(\beta) + \sqrt{n}(\hat{\beta}_1 - \beta_1)\frac{\partial}{\partial \beta_1}h(\beta) + o_p(\sqrt{n}).$$

By Slutsky's theorem, the remainder term can be ignored. For any 2-vector v,

$$\sqrt{n}v^\top(\hat{\beta} - \beta) \to_D N_1(0, v^\top I^{-1} v),$$

so for

$$v^\top = (\frac{\partial}{\partial \beta_0}h(\beta), \frac{\partial}{\partial \beta_1}h(\beta)),$$

the finite sample approximation is

$$\begin{aligned}(\widehat{ED}\alpha - ED\alpha) \quad &\equiv_D \quad v^\top(\hat{\beta} - \beta) \\ &\equiv_D \quad N(0, (\frac{\partial}{\partial \beta_0}h)^2 \,\rho_{11} + 2(\frac{\partial}{\partial \beta_0}h)(\frac{\partial}{\partial \beta_1}h)\,\rho_{12} + (\frac{\partial}{\partial \beta_1}h)^2\,\rho_{22}),\end{aligned}$$

where $(\rho_{kl})_{1 \le k,l \le 2} = I^{-1}/n$. We obtain consistent estimates for these elements from the inverse of the finite Fisher information matrix

$$(X^\top W X)^{-1} = (\hat{\rho}_{kl})_{1 \le k,l \le 2},$$

a matrix that is available from output of programs. Substituting the estimated covariance matrix is permitted by a second application of Slutsky's theorem.

Specifically, for the dose-response model without log transformation,

$$h(\beta) = \frac{F^{-1}(\tilde{\alpha}) - \beta_0}{\beta_1},$$

$$\frac{\partial h}{\partial \beta_0} = -\frac{1}{\beta_1},$$

$$\frac{\partial h}{\partial \beta_1} = -\frac{F^{-1}(\tilde{\alpha}) - \beta_0}{\beta_1^2}$$

so that an approximate $100(1-\gamma)\%$ c.i. for $ED\alpha$ is given by

$$\widehat{ED}\alpha \pm \Phi^{-1}(1-\frac{\gamma}{2})\left[\frac{\hat{\rho}_{11}}{\hat{\beta}_1^2} + 2\frac{F^{-1}(\tilde{\alpha})-\hat{\beta}_0}{\hat{\beta}_1^3}\hat{\rho}_{12} + (\frac{F^{-1}(\tilde{\alpha})-\hat{\beta}_0}{\hat{\beta}_1^2})^2\hat{\rho}_{22}\right]^{1/2}.$$

Other methods to obtain inference for $ED\alpha$ have been developed (Fieller, Spearman-Kärber, Robbins-Monro, up-and-down-method, smoothing methods). Important is the estimation of $ED\alpha$ for small α (low-dose extrapolation problem), for both toxicology and hormesis.
For confidence intervals, one can also use the bootstrap (various bootstraps have been developed and compared for this application in the literature).

18.6 Comparison of dose-response functions

Response quotient function

Compare dose-response for two substances A, B assuming log dose scales:

$$p_A(d) = F_A(\beta_{0A} + \beta_{1A}\log d)$$

$$p_B(d) = F_B(\beta_{0B} + \beta_{1B}\log d)$$

Response quotient function (equating α and $\tilde{\alpha}$)

$$q(\alpha) = \frac{p_B^{-1}(\alpha)}{p_A^{-1}(\alpha)} = \frac{ED_B\alpha}{ED_A\alpha}$$

corresponds to relative strength of B as compared to A to achieve the same effect of size α.

Application: Effectiveness of biologically produced vaccines against a proven standard. If $F_A = F_B = F$ (identical link functions), then

$$q(\alpha) = \exp\{(F^{-1}(\alpha) - \beta_{0B})/\beta_{1B} - (F^{-1}(\alpha) - \beta_{0A})/\beta_{1A}\}$$

Logparallel assays

If $\beta_{1A} = \beta_{1B} = \beta_1$ (the so-called assumption of *log-parallel slopes*), then we have a logparallel assay:

$$q(\alpha) = \exp(\frac{\beta_{0A} - \beta_{0B}}{\beta_1}) \equiv const. = q,$$

indicating the *relative strength* of A vs. B.

Example: Dilutions lead to logparallel assays, i.e., if a substance B has the same effect as a diluted version of substance A with dilution factor ζ, then the dose d_B of substance B corresponds to that of dose ζd_A of substance A and therefore

$$q(\alpha) = q = \zeta.$$

Testing for log-parallelity:

Fit the 4 parameter model with linear predictor

$$\begin{aligned} \eta &= \beta_{0A} + (\beta_{0B} - \beta_{0A})\, 1_{\{\text{substance}B\}} + \beta_{1A} \log d + (\beta_{1B} - \beta_{1A}) \log d\, 1_{\{\text{substance}B\}} \\ &= \alpha_1 + \alpha_2\, 1_{\{B\}} + \alpha_3 \log d + \alpha_4 \log d\, 1_{\{B\}} \end{aligned}$$

to the combined data, with joint link function.

Logparallelity is equivalent to

$$H_0: \quad \alpha_4 = 0.$$

This is a standard Wald test based on asymptotic normality. To accept, the general strategy is to test at level $\alpha = 0.1$ and if the test does not reject at this level this will be interpreted as acceptance.

If we accept H_0, then $\alpha_4 = 0$ which means we work with the model that has only parameters

$\alpha_1, \alpha_2, \alpha_3$ and $p = 3$. Fitting the parameter estimates within this reduced model,

$$\hat{q} = \exp(-\hat{\alpha}_2/\hat{\alpha}_3)$$

is the estimate for relative strength.

Inference for q is obtained by another application of the δ-method, that is analogous to the above application for the $ED\alpha$. The starting point is the result

$$\sqrt{n}(\hat{\alpha} - \alpha) \to_D N_3(0, I^{-1}), \quad \alpha = (\alpha_1, \alpha_2, \alpha_3)^\top,$$

with finite sample approximation $\hat{\alpha} \sim N_3(\alpha, (X^\top W X)^{-1})$. Writing similarly as above

$$(X^\top W X)^{-1} = (\hat{\rho}_{kl})_{1 \le k,l \le 3},$$

one obtains, applying the δ-method to $\log \hat{q} = -\hat{\alpha}_2/\hat{\alpha}_3$, in analogy to the arguments for the $ED\alpha$, the asymptotic $100(1-\gamma)\%$ c.i. for the relative strength q as follows,

$$[\hat{q} \exp\left(-\Phi^{-1}(1 - \frac{\gamma}{2})\, \hat{\rho}^{1/2}\right), \ \hat{q} \exp\left(\Phi^{-1}(1 - \frac{\gamma}{2})\, \hat{\rho}^{1/2}\right)], \quad \text{with } \hat{\rho} = \frac{\hat{\rho}_{22}}{\hat{\alpha}_3^2} - 2\frac{\hat{\alpha}_2}{\hat{\alpha}_3^3}\hat{\rho}_{23} + \frac{\hat{\alpha}_2^2}{\hat{\alpha}_3^4}\hat{\rho}_{33} \tag{18.1}$$

19 Epidemiological Study Designs

Causation of disease: Risk factors, Exposure factors. Cannot control variables as in dose-response studies, since data are observational.

Example: Risk factor → Disease

Alcohol consumption → liver cirrhosis

Study designs: Cohort study, Case-control study.

19.1 Cohort Studies

Prospective study:

Example: Framingham study for coronary heart disease, 1948-1990, 2187 men and 2669 women. Start with disease-free individuals and follow them for a long period of time.

Historical cohort study: go back in time to assess Risk factors/Exposure (problem: incomplete records).

Occupational studies: "Person years of exposure" as predictor variable, cross-classified with other risk factors.

Example: Chemical workers ("healthy worker effect"), intermittent exposure by divers (Caisson disease).

Incident Cohorts: Patients are enrolled at time of diagnosis/first treatment.

Prevalent Cohorts: Patients enrolled some time after diagnosis/first treatment. Time from diagnosis/first treatment to enrollment usually will be an important covariate.

Cohort studies often allow a survival analysis approach: Common *endpoints* are "Time to Disease", e.g., as a function of person-years of exposure, or "Time to Death" and "Time to Hospitalization". Censored and missing data is often a major problem that needs to be addressed.

Inference for time-varying covariates: $X(t)$ is a covariate process, for example, person-years of exposure up to time t (increasing), or marker of a disease process such as drug prescriptions (driven by seriousness of disease), cholesterol level, or cancer biomarkers. An option is to fit a Cox proportional hazards model,

$$\lambda(t) = \lambda_0(t) \exp\{\beta' X(t) + \gamma Z\}$$

where $X(t) = (X_1(t), \ldots, X_p(t))^\top$ are one or several such time-dependent exposures, and Z are baseline measurements such as patient characteristics; here $\lambda(\cdot)$ is the hazard function, $\lambda_0(\cdot)$ the baseline hazard.

19.2 Case-control studies

Cases: Have the disease of interest

Controls: Do not

Compare exposure to risk factors between these groups; it is a retrospective study, often fraught by bias problems

In the simplest cases, data are arranged in a 2×2 table:

Example: Diabetes as risk factor for cataract.

	Cataract Cases	Controls
Diabetes	55	84
Non-Diabetes	552	1927

Case-control studies are cheaper but much less reliable than cohort studies; exposure assessment is often unreliable. Cases and controls must be representative of the populations.

In a cohort study, one can determine *risk* for a diabetic to develop cataract; in a case-control study only *relative risk*.

19.3 Measures of risk

Basic problem: Which time frame do we consider for disease to be observed? Usually, longitudinal survival analysis approaches much preferred. Risk is then measured via the *hazard ratio* (HR) in the proportional hazards model (assuming the proportional hazards assumption is reasonable).

Let p_e, p_u be the risk (probability) to get the disease for exposed and unexposed people. *Relative risk* of exposed versus non-exposed persons is

$$\rho = \frac{p_e}{p_u}.$$

Odds Ratio: Odds for disease in exposed/unexposed groups are $\frac{p_e}{1-p_e}$, $\frac{p_u}{1-p_u}$, and odds ratio is

$$\psi = \frac{p_e/(1-p_e)}{p_u/(1-p_u)}$$

Note for small p_e, as is usually assumed,

$$\frac{p_e}{1-p_e} = p_e(1 + p_e + p_e^2 + \cdots \approx p_e,$$

and similarly for p_u, so that for small risks

$$\rho \approx \psi.$$

Sample 2 × 2 table (with data from a random sample of population)

	D	$\overline{D}$
E	a	b
$\overline{E}$	c	d

Notation: E exposed, $\overline{E}$ not exposed, D diseased, $\overline{D}$ not diseased.

Note

$$\hat{p}_e = \frac{a}{a+b}, \;\; \hat{p}_u = \frac{c}{c+d}, \;\; \hat{\rho} = \frac{\hat{p}_e}{\hat{p}_u},$$

$$\hat{\psi} = \frac{\hat{p}_e/(1-\hat{p}_e)}{\hat{p}_u/(1-\hat{p}_u)} = \frac{ad}{bc}.$$

Can show:

$$\log(\hat{\psi}) \approx N\left(\log(\psi), \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)$$

for large sample size. Then

$$\log\hat{\psi} \pm \Phi^{-1}(1 - \frac{\alpha}{2})\hat{\sigma}$$

is approximate $100(1-\alpha)\%$ c.i. for $\log(\psi)$, where

$$\hat{\sigma}^2 = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d},$$

and

$$[\hat{\psi}\exp\{-\Phi^{-1}(1-\frac{\alpha}{2})\hat{\sigma}\}, \hat{\psi}\exp\{\Phi^{-1}(1-\frac{\alpha}{2})\hat{\sigma}\}]$$

is $100(1-\alpha)\%$ c.i. for ψ.

19.4 Confounding

Exposure is associated with an additional risk factor (confounding factor). Ignoring this factor leads to uninterpretable odds ratio estimates.

Example: Coronary heart disease (CHD), Risk factor: Alcohol

Alcohol	cases	controls
high	68	33
low	32	72

$$\hat{\psi} = 4.64$$

Alcohol in fact is known to be unrelated to CHD, or has an odds ratio $\psi < 1$. In the above analysis, smoking is ignored, while smoking is closely associated with alcohol consumption (smokers are more likely to drink alcohol), and is a real risk factor for CHD. It is important to include all applicable risk factors in a model in order to avoid confounding. A confounding variable can modify the effect of a risk factor (*effect modifier*), a situation that can be modelled through an interaction term.

20 Logistic Regression for Cohort Studies

Alternative to survival analysis approach (the survival approach is usually preferred. Whenever feasible, it has more power).

Data: $y_i, X_{1i}, X_{2i}, \dots, X_{ki},\ \ i = 1, 2, \dots, n.$

Outcome y_i =Survival Time, $\delta_i = \begin{cases} 0 & \text{obs censored} \\ 1 & \text{uncensored} \end{cases} \Rightarrow$ Cox PH model

$$\text{or } y_i = \begin{cases} 1 & \text{subject develops disease during follow-up period} \\ 0 & \text{no} \end{cases}$$

$$p_i = E(y_i|X_i), \quad \text{logit}(p_i) = \log \frac{p_i}{1-p_i} = \eta_i \text{ (linear predictor)}$$

Principles: Include all confounding variables, and all possibly relevant exposure/risk factors (Important is to adjust risk factors for confounders). Especially, consider interactions between exposure factors and between exposure factors and confounders.

Example: One discrete exposure factor x as predictor (observed at levels $x = 0$ – unexposed; $x = 1$ – exposed)

$$\text{logit}(p) = \beta_0 + \beta_1 x.$$

Setting p_u = probability of disease for unexposed person, p_e = for exposed person, one obtains:

Odds of disease for unexposed person:

$$\frac{p_u}{1-p_u} = e^{\beta_0}.$$

Odds for exposed person:

$$\frac{p_e}{1-p_e} = e^{\beta_0+\beta_1}.$$

Odds ratio:

$$\psi = \frac{p_e/(1-p_e)}{p_u/(1-p_u)} = e^{\beta_1}, \quad \beta_1 = \log\psi, \quad \hat{\psi} = e^{\hat{\beta}_1},$$

and $(1-\alpha)$ c.i. for ψ is

$$[e^{\hat{\beta}_1 - \Phi^{-1}(1-\frac{\alpha}{2})s(\hat{\beta}_1)}, \; e^{\hat{\beta}_1 + \Phi^{-1}(1-\frac{\alpha}{2})s(\hat{\beta}_1)}]$$

Can easily extend this to an exposure variable with m levels; with $(m-1)$ dummy variables, one can define odds ratios ψ_{kl} between exposure at level l and at level k,

$$\hat{\psi}_{kl} = e^{\hat{\beta}_{k+1} + \cdots + \hat{\beta}_l}.$$

Continuous exposure variable:

$$\operatorname{logit}(p) = \beta_0 + \beta_1 x, \qquad x \text{ continuous}$$

$$\psi = \frac{\exp(\beta_0 + \beta_1(x+1))}{\exp(\beta_0 + \beta_1 x)} = e^{\beta_1}.$$

This is the odds ratio between risk when exposure variable is increased by 1, i.e., at $x+1$ and the basic level of the exposure variable, i.e., x.

Example: Increase systolic blood pressure from 150 to 200 mm Hg and observe the resulting increase in risk of stroke.

Note: The odds ratio does not depend on the level x. But it does depend on the size of the units in which the exposure variable x is measured. Using larger units increases the odds ratio. This can lead to misleading presentation of numbers. Significance tests are not affected by the change in units.

21 Logistic Regression for Case-Control Studies

Define sampling fraction of cases

$$\pi_1 = P(\text{individual is in study} | \text{ diseased}) = P(S|D)$$

and sampling fraction of controls

$$\pi_2 = P(\text{individual is in study } | \text{ disease-free}) = P(S|\bar{D});$$

$\pi_1, \ \pi_2$ are unknown and cannot be estimated in a case-control study. Important assumption: S and the covariates are independent. This is a strong assumption that is often violated.

Example: Screening for breast cancer: If age is a covariate, older women with cancer are more likely to be in the study due to screening which takes place at certain ages. The assumption is violated.

Of interest:

$$\begin{aligned} p(x) &= P\text{ (individual has disease when covariates at level } x) \\ &= P(D|X = x) \end{aligned}$$

where the covariates consist of confounders and risk factors.
By Bayes theorem, and using the above assumptions, we may express the probability for a case in the case-control study as follows. For the fourth equation below use that
$S = \{\text{Subject is in the study sample}\}$ and x are independent:

$$\begin{aligned} p_0(x) &= P(D|X = x, \text{subject in study}) \\ &= \frac{P(S, D, x)}{P(S, D, x) + P(S, \bar{D}, x)} \\ &= \frac{P(D, x)P(S|D, x)}{P(D, x)P(S|D, x) + P(\bar{D}, x)P(S|\bar{D}, x)} \\ &= \frac{P(D|x)P(S|D)}{P(D|x)P(S|D) + (1 - P(D|x))P(S|\bar{D})} \\ &= \frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_2(1 - p(x))}. \end{aligned}$$

Then

$$\frac{p_0(x)}{1 - p_0(x)} = \frac{\pi_1}{\pi_2} \frac{p(x)}{1 - p(x)}, \tag{21.1}$$

$$\mathrm{logit}(p_0(x)) = \log(\frac{\pi_1}{\pi_2}) + \mathrm{logit}(p(x))$$

Assume now that

$$\mathrm{logit}(p(x)) = \eta \quad \text{(linear predictor), then}$$

$$\begin{aligned}\mathrm{logit}(p_0(x)) &= \log(\frac{\pi_1}{\pi_2}) + \beta_0 + \sum_{l=1}^{p-1} \beta_l x_l \\ &= \tilde{\beta}_0 + \sum_{l=1}^{p-1} \beta_l x_l.\end{aligned}$$

Data: Binary response variable $y_i = \begin{cases} 1 & \text{case} \\ 0 & \text{control} \end{cases}$, covariates are $x_{i1}, \ldots, x_{ip-1}$.

Here, β_0 is not estimable, therefore $p(x)$ is not estimable.

Considering

$$\psi_{x_1 x_0} = \frac{p(x_1)/(1-p(x_1))}{p(x_0)/(1-p(x_0))},$$

the odds ratio for those at exposure levels x_0 and x_1, (21.1) gives

$$\frac{p(x_j)}{1-p(x_j)} = \frac{\pi_2}{\pi_1} \frac{p_0(x_j)}{1-p_0(x_j)}, \quad j = 0, 1,$$

therefore

$$\psi_{x_1 x_0} = \frac{p_0(x_1)/(1-p_0(x_1))}{p_0(x_0)/(1-p_0(x_0))} = \frac{p(x_1)/(1-p(x_1))}{p(x_0)/(1-p(x_0))},$$

which does not depend on β_0, and can be estimated via

$$\log \hat{\psi}_{x_1 x_0} = \mathrm{logit}(\hat{p}_0(x_1)) - \mathrm{logit}(\hat{p}_0(x_0)) = \sum_{l=1}^{p-1} \hat{\beta}_l (x_1 - x_0),$$

$$\hat{\psi}_{x_1 x_0} = \exp(\sum_{l=1}^{p-1} \hat{\beta}_l (x_1 - x_0)).$$

Again, this provides a rationale for using odds ratios in epidemiological risk assessment.

22 Matched Case-Control Studies

Adjustment for confounders is a central issue in epidemiological data analysis. Can include confounding variables as covariates into the model, as above, or can control for confounding by *matching*.

Matched case-control study: For each case, we assign one or several *matched controls*: Matched controls should be as close as possible to case in regard to the values of confounding variables.
Typical confounders: Sex, age, ethnic group, smoker status, etc..
A design with M controls is a 1:M matched study.

Example:

	Case	controls
Outcome:	Accidental death in the house	no death

Exposure/Risk: gun ownership yes or no
Confounders: State of residence, household income, family size, domestic violence (if available)
Most common: 1:1 matching

Assume there are n ”matched sets” in a study with $1 : M$ matching:
The j-th matched set is ($j = 1, \ldots, n$):

disease status	$y_{0j} = 1$	$y_{1j} = 0,$	$\cdots,$	$y_{Mj} = 0$
Covariates $\in \mathcal{R}^{p-1}$	$x_{0j},$	$x_{1j},$	$\cdots,$	x_{Mj}
	case		M controls	

Let D=”diseased”=case, $\overline{D}$=”not diseased”=control, f be a pdf or pmf, as appropriate, and $f(x|D)$, $f(x|\overline{D})$ be the conditional pmf/pdf’s.

Conditional likelihood for the j-th matched set:

$$\begin{aligned} L_j &= \frac{f(x_{0j},\ldots,x_{Mj}|y_{0j}=1,y_{1j}=0=\ldots=y_{Mj}=0)}{f(x_{0j},\ldots,x_{Mj}|\text{ One case, M controls})} \\ &= \frac{f(x_{0j}|D)\prod_{i=1}^{M} f(x_{ij}|\overline{D})}{\sum_{i=0}^{M} f(x_{ij}|D)\prod_{l=0,l\neq i}^{M} f(x_{lj}|\overline{D})} \\ \text{(Bayes)} &= \frac{f(D|x_{0j})\prod_{i=1}^{M} f(\overline{D}|x_{ij})}{\sum_{i=0}^{M} f(D|x_{ij})\prod_{l=0,l\neq i}^{M} f(\overline{D}|x_{lj})} \\ &= \Big[1+\frac{\sum_{i=1}^{M} f(D|x_{ij})\prod_{l=0,l\neq i}^{M} f(\overline{D}|x_{lj})}{f(D|x_{0j})\prod_{i=1}^{M} f(\overline{D}|x_{ij})}\Big]^{-1} \\ &= \Big[1+\sum_{i=1}^{M}\frac{f(D|x_{ij})}{f(D|x_{0j})}\,\frac{f(\overline{D}|x_{0j})}{f(\overline{D}|x_{ij})}\Big]^{-1}, \end{aligned}$$

using $a/b = 1/(1+(b-a)/a)$. Conditional likelihood for all matched sets: $L=\prod_{j=1}^{n} L_j$. Assume logit link:

$$P(D|x_{ij}) = f(D|x_{ij}) = \text{expit}\{\eta_{ij}\}, \quad \eta_{ij} = \alpha_j + \sum_{k=1}^{p-1}\beta_k x_{ijk}$$

where $x_{ij} = (x_{ij1},\ldots,x_{ijp-1})^\top$:

$$\frac{P(D|x_{ij})}{P(\overline{D}|x_{ij})} = \frac{e^{\eta_{ij}}/(1+e^{\eta_{ij}})}{1-e^{\eta_{ij}}/(1+e^{\eta_{ij}})} = e^{\eta_{ij}}$$

$$\Rightarrow L = \prod_{j=1}^{n}\Big[1+\sum_{i=1}^{M} e^{\eta_{ij}-\eta_{0j}}\Big]^{-1} = \prod_{j=1}^{n}\Big[1+\sum_{i=1}^{M}\exp\{\sum_{k=1}^{p-1}\beta_k(x_{ijk}-x_{0jk})\}\Big]^{-1}$$

We obtain the conditional likelihood estimates

$$\hat{\beta} = \text{argmax}_\beta L(\beta)$$

with the usual interpretation

$$\log\hat{\psi}_k = \hat{\beta}_k, \;\; \hat{\psi}_k = e^{\hat{\beta}_k}$$

for the odds ratio of the k-th risk factor. We cannot estimate the intercepts α_j, which are specific effects for the j-th matched set.

This approach is called conditional logistic regression. The usual asymptotics for a likelihood apply, in terms of asymptotic normal distributions and inference.

A computational trick is possible for the common 1:1 matched studies: Here,

$$L = \prod_{j=1}^{n} [\, 1 + \exp\{\sum_{k=1}^{p-1} \beta_k (x_{1jk} - x_{0jk})\}\,]^{-1},$$

as $M = 1$. Defining $z_{jk} = x_{1jk} - x_{0jk}, \quad \eta_j = \sum_{k=1}^{p-1} \beta_k z_{jk}$,

$$\begin{aligned} L &= \prod_{j=1}^{n} [\, 1 + \exp\{\sum_{k=1}^{p-1} \beta_k z_{jk}\}\,]^{-1} \\ &= \prod_{j=1}^{n} [\, 1 - \frac{e^{\eta_j}}{1 + e^{\eta_j}}\,] \\ &= \prod_{j=1}^{n} (1 - p_j)^{1-y_j} p_j^{y_j}, \end{aligned}$$

where

$$p_j = \frac{e^{\eta_j}}{1 + e^{\eta_j}}, \;\; y_j = 0, \;\; j = 1, \ldots, n.$$

This is the likelihood for binomial regression with logit link, outcomes $y_j = 0$, and linear predictor $\eta_j = \sum_{k=1}^{p-1} \beta_k z_{jk}$ without intercept. This model can thus be fitted with common GLM programs, using these specifications.

23 Diagnostic Tests

Based on clinical measurements (outcomes of tests, blood measurements, blood pressure, etc.) the disease status is to be predicted (*diagnosis*). We fit a binomial regression model to data from a *gold standard*: Predictors and true disease status are known.

Data: Predictors X_i, Diagnosis $y_i = \begin{cases} 1 & \text{yes} \\ 0 & \text{no} \end{cases}$

$$p_i = P(y_i = 1|X_i) = g^{-1}(\eta_i), \quad \eta_i = X_i\beta.$$

The common default choice is the logit link.

Diagnostic test: If $p_i > p_0$, one would declare a positive diagnosis, i.e., presence of the disease. Here, p_0 serves as a cut-point; if $p_i \leq p_0$, disease is not diagnosed. If D=disease, T+=Test outcome is positive, then

$$P(T + |D) = \textit{sensitivity}, P(T - |\overline{D}) = \textit{ specificity},$$

$$P(D|T+) = \textit{positive predictive value}, \;\; P(\overline{D}|T-) = \textit{ negative predictive value}$$

of the diagnostic test.
By Bayes' theorem, one obtains the basic formula

$$P(D|T+) = \frac{P(T + |D)P(D)}{P(T + |D)P(D) + (1 - P(T - |\overline{D}))(1 - P(D))},$$

demonstrating that positive predictive value is determined by sensitivity, specificity and prevalence.

Let $f(x|D)$, $f(x|\overline{D})$ be the densities of the distributions of the predictors, conditional on disease status (diagnosis) and F_D, $F_{\overline{D}}$ the corresponding cdfs. Noting $p > p_0$ is equivalent to $\eta > \eta_0$ for a unique η_0, sensitivity and specificity as functions of the threshold η_0 become:

$$\text{Sensitivity}(\eta_0) = \int_{\eta > \eta_0} f(x|D)dx = 1 - F_D(\eta_0),$$

$$\text{Specificity}(\eta_0) = \int_{\eta \leq \eta_0} f(x|\overline{D})dx = F_{\overline{D}}(\eta_0).$$

Receiver-Operator Characteristic (ROC)

The goal is to choose η_0 in an optimal way. The ROC curve is the graph

$$(1 - \text{Specificity}(\eta_0),\;\; \text{Sensitivity}(\eta_0)) \equiv (1 - F_{\overline{D}}(\eta_0),\, 1 - F_D(\eta_0)),$$

implicity parametrized in η_0. Equivalently,

$$\text{ROC}(u) = 1 - F_D(F_{\overline{D}}^{-1}(1-u)), \quad 0 \le u \le 1.$$

Comparing diagnostic tests

Note that in the function ROC(u), the x-axis u corresponds to specificity and the y-value $1 - F_D(F_{\overline{D}}^{-1}(1-u))$ to the corresponding sensitivity, when choosing threshold $\eta_0(u) = F_{\overline{D}}(1-u)$. Since by varying η_0, all specificities $0 \le u \le 1$ will be covered, a test is better where in general the corresponding Sensitivity(u) =ROC(u) is larger.

Therefore: A criterion for comparing tests is the area under the curve ROC(u), i.e., $\int_0^1 \text{ROC}(u)\, du$, which is larger for a better diagnostic test and should be maximized. For the important choice of η_0 (the cut-point), one can use criteria such as $\hat{\eta}_0 = \text{argmax}_\eta\{Spec.(\eta_0)^2 + Sens.(\eta_0)^2\}$.

In order to implement these criteria, one needs to estimate $F_D, F_{\overline{D}}$ from training data and plug in. To do this requires a *gold standard* where diagnosis is 100% accurate.

24 GLMs for correlated/longitudinal data

This is a direct extension of the linear random effects model to the case of clustered data, that arise in longitudinal studies and repeated measurements.

The repeated measurements for the i-th subject or cluster and associated responses are given by

$$\frac{x_{i1}}{y_{i1}}, \frac{x_{i2}}{y_{i2}}, \ldots, \frac{x_{im_i}}{y_{im_i}},$$

where predictors $x_{ij} \in \mathbb{R}^{p-1}$ and responses $y_{ij} \in \mathbb{R}$, and the subject indices are $i = 1, \ldots, n$ for a total of n subjects or clusters. The number of measurements made for the i-th subject or cluster is m_i.

24.1 Working correlations

The correlation structure of the responses can either depend on a parameter that can be included in likelihood or quasi-likelihood, or it can be one of a number of common working correlation structures:
These are defined by means of $m_i \times m_i$ correlation matrices R_i:
Unspecified: General form of $R_i = (\rho_{kl})_{1 \le k \le m_i, 1 \le l \le m_i}$ with arbitrary $0 \le \rho_{kl} \le 1, \rho_{kl} = \rho_{lk}$ and $\rho_{kk} = 1$, furthermore R_i are non-negative definite.

Exchangable/Compound Symmetric: $R_i = (\rho_{kl})_{1 \le k \le m_i, 1 \le l \le m_i}$ with $\rho_{kk} = 1$ and $\rho_{kl} = \rho, k \neq l$, for a constant $0 \le \rho \le 1$.

Independent: $R_i = I_{m_i \times m_i}$

Autoregressive: $R_i = (\rho_{kl})_{1 \le k \le m_i, 1 \le l \le m_i}$ with $\rho_{kl} = \rho^{|k-l|}$ for a given ρ with $0 \le \rho \le 1$.

Example: Repeated measurements model:

$$Y_{ij} = \xi_i + \varepsilon_{ij}$$

with random subject effects ξ_i and measurement errors ε_{ij}, such that $1 \le i \le n, 1 \le j \le m_i$ and $E(\xi_i) = C_i, \operatorname{var}(\xi_i) = \sigma_\xi^2$ for constants C_i, and $E(\varepsilon_{ij}) = 0, \operatorname{var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$, where all ε_{ij} are i.i.d., the ξ_i are i.i.d., and the ε_{ij} and the ξ_i are independent.

Then

$$\mathrm{cov}(Y_{ij}, Y_{ik}) = \mathrm{cov}(\xi_i + \varepsilon_{ij}, \xi_i + \varepsilon_{ik}) = \begin{cases} \sigma_\xi^2 & j \neq k \\ \sigma_\xi^2 + \sigma_\varepsilon^2 & j = k \end{cases}$$

Thus this gives compound symmetry with

$$\rho = \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\varepsilon^2}.$$

24.2 Generalized Estimating Equations (GEE)

Given are $m_i \times m_i$ working correlation matrices R_i and a variance function $V(\cdot)$. Define $m_i \times m_i$ matrices

$$Q_i = diag(V(\mu_{i1}), \ldots, V(\mu_{im_i})), \quad \text{and} \quad V_i = R_i Q_i.$$

Let $N = \sum_{i=1}^n m_i$ be the total no. of measurements and define the $N \times p$ matrix

$$D = \left(\frac{\partial \mu_l}{\partial \beta_j} \right), \quad l = 1, \ldots, m_1, m_1 + 1, \ldots, m_1 + m_2, \ldots, \sum_{i=1}^n m_i, \quad j = 0, \ldots, p-1.$$

Combine the matrices $m_i \times m_i$ matrices V_i into a $N \times N$ block matrix

$$V = \begin{bmatrix} V_1 & 0 & \ldots & 0 \\ 0 & V_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & V_n \end{bmatrix}$$

Then obtain generalized estimating equations (GEE) in analogy to estimating equations for quasi-likelihood,

$$U(\beta) = \begin{pmatrix} U_1(\beta) \\ \vdots \\ U_p(\beta) \end{pmatrix} = D^\top V^{-1}(y - \mu)/\sigma^2 = 0,$$

and

$$\mathrm{cov}(U(\beta)) = D^\top V^{-1} D/\sigma^2 = I_{GEE},$$

analogous to the derivation of quasi-information.

Assuming

$$(\frac{D^\top V^{-1} D}{n})^{-1} \to_{n\to\infty} \Sigma,$$

one may obtain

$$\sqrt{n}(\hat{\beta} - \beta) \to_D N(0, \sigma^2 \Sigma),$$

which can be used for inference.

Note: The derivation is more complex than in the quasi-likelihood case. The working correlation type needs to be specified and usually there are unknown parameters that are obtained via a second estimating equation. This GEE approach is a *marginal* or *population average* (PA) approach since only the second moments of the observations matter, and no subject random effects are estimated.

24.3 Random effects models

Mixed linear model:

$$Y = X\beta + Z\gamma,$$

where X are the predictors for fixed effects (combined into a design matrix), Z are the predictors for the random effects (often chosen as $Z = X$), and β, γ are fixed parameters respectively random effects. The random effects assume subject-specific values.

Generalized Linear Mixed Model (GLMM): For a link function g and linear predictor η,

$$g(E(Y|X, Z)) = \eta, \quad \eta = X\beta + Z\gamma,$$

where we assume

$$\gamma \sim F,$$

and the distribution F is customarily assumed multivariate normal.

These models are *subject-specific*, as every subject is modelled via a random effect. Typical assumption is that given random effects γ_i, responses $Y_{i1}, Y_{i2}, \ldots, Y_{im_i}$ are independent and follow a GLM.

Marginalizing a GLMM:

$$\begin{aligned} E[E(Y|X,Z,\gamma)|X,Z] &= E[E(Y|X,Z,\gamma)|X,Z] \\ &= E\{[g^{-1}(X\beta + Z\gamma)]|X,Z\} \\ &= \int g^{-1}(X\beta + Zu)dF(u). \end{aligned}$$

In general, this model cannot be explicitly represented by GEE with known link. This integration can be very complicated, and only under special circumstances will one have a known link function h s.t. $E(Y|X) = h^{-1}(X\beta)$ as in the usual GLM.

Likelihood:

$$L(\beta; Y) = \prod_{i=1}^{n} \int \prod_{j=1}^{m_i} f(y_{ij}|u)\, dF(u),$$

where $f(\cdot|u)$ is the conditional pdf of Y_{ij} given γ_i. The evaluation of this likelihood generally requires a difficult integration step.

25 GLMs with Functional Predictors

Functional predictors consist of time courses or profiles that serve as predictor variables, viewed as (infinite-dimensional) random functions or trajectories. Random trajectories = realizations of a stochastic process.

Examples: Gene expression time courses; pharmacokinetic distribution curves; CD4 cell count trajectories for HIV-infected subjects. Responses include continuous (remaining lifetime) or binary (diagnosis) observations.

Infinite dimension of trajectories requires dimension reduction and regularization. A common tool for dimension reduction is principal component analysis, which can be extended to the case of random functions.

25.1 Principal Component Analysis for random functions

Let $G : [0,T]^2 \to \mathbb{R}$ be a continuous symmetric nonnegative function which is also non-negative definite, i.e., for all functions g, $\int\int G(s,t)g(s)g(t)dtds \geq 0$.
Define a linear integral operator A_G on $L^2([0,T])$ by $(A_G X)(s) = \int_0^T G(s,t)X(t)dt$, for all $X \in L^2([0,T])$, $A_G : L^2([0,T]) \to L^2([0,T])$. Eigenvalues λ and eigenfunctions ϕ of the (non-neg. definite) operator A_G are defined by

$$(A_G\phi)(s) = \lambda\phi(s), \quad s \in [0,T], \quad \lambda \in \mathbb{R}.$$

The set of all eigenvalues constitute the spectrum of A_G. There are countably many eigenvalues which are all real-valued non-negative, with a limit point at 0.

Let $\{\phi_j\}_{j=1,2,\dots}$ be an orthonormal basis for the space corresponding to the eigenfunctions of A_G, with corresponding eigenvalues $(\lambda_j)_{j=1,2,\dots}$, $\lambda_1 \geq \lambda_2 \geq \dots$.

Mercer's theorem: $G(s,t) = \sum_{j=1}^\infty \lambda_j\phi_j(s)\phi_j(t)$, i.e., the operator A_G is determined by its eigenfunctions/eigenvalues.
Note:
1. G is called the kernel of the operator.
2. An operator which can be represented as a linear integral operator with positive definite symmetric kernel is a Hilbert-Schmidt operator.

For a random curve $X(t) \in L^2([0,T])$, i.e., $E\int X^2(t)dt < \infty$, and also $\sup_t E(X^2(t)) < \infty$, assume

$$\begin{aligned} EX(t) &= \mu(t) \quad \text{(mean function)} \\ \operatorname{cov}(X(t), X(s)) &= E(X(t)-\mu(t))(X(s)-\mu(s)) = G(s,t) \quad \text{(covariance function)} \end{aligned}$$

is continuous in s and t.

If the operator A_G has eigenfunctions/eigenvalues $(\phi_j, \lambda_j)_{j=1,2,\ldots}$, then X has the following representation.

Karhunen-Loève Theorem:

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j \phi_j(t), \quad 0 \le t \le T,$$

where $\xi_j = \int_0^T (X(t)-\mu(t))\phi_j(t)dt$ (inner products $= \langle X-\mu, \phi_j \rangle$) are r.v.'s with $E\xi_j = 0$, $E\xi_j\xi_k = 0$ (uncorrelated) for $j \neq k$, and $E\xi_j^2 = \lambda_j$, where $\sum \lambda_j \le \infty$. It holds that $E[X(t) - \sum_{j=1}^{M} \xi_j\phi_j(t)]^2 \to 0$ uniformly in $t \in [0,T]$, as $M \to \infty$.
The ξ_j are the *Functional Principal Components* (FPCs) of X.

Example: Let $G(s,t) = \min(s,t)$, and assume $X(\cdot)$ are Gaussian random functions. All joint distribution $(X(s_1), \ldots, X(s_m)) \sim$ Gaussian. Then $X(\cdot)$ is the Brownian motion (continuous but no-where differentiable) with $X(s) \sim N(0,s)$ and independent increments;
$\operatorname{var}(X(t)) = t$; $\operatorname{cov}(X(t), X(s)) = \operatorname{cov}(X(s), X(s) + (X(t) - X(s))) = \operatorname{var}(X(s)) = s = \min(s,t)$.
Need to solve eigen-equation

$$\int \min(t,s)\phi(t)dt = \lambda\phi(s).$$

Solution: $\phi_k(s) = \sqrt{2}\sin\{(2k-1)\frac{\pi s}{2}\}$, λ_k complicated expression. For $\xi_j \sim N(0,1)$ (in this case, scores are independent, not just uncorrelated), i.i.d, let $\xi_j^* = \xi_j\sqrt{\lambda_j} \Rightarrow$

$$X(t) = \sqrt{2}\sum_{j=1}^{\infty} \xi_j^* \sin\{(2j-1)\frac{\pi t}{2}\}$$

(Representation of Brownian motion in principal components).

25.2 Generalized Functional Linear Model

Predictors $X_i \in L^2$, associated generalized responses Y_I (independent r.v.s, Poisson, Binomial, ...), $i = 1, \ldots, n$.

We use a QL approach for this generalized regression problem: Let $\beta(\cdot) \in L^2$ be a parameter function;
$g(\cdot)$ a suitable link function (adapted to responses);
$\sigma^2(\cdot)$ a variance function, e.g., $\sigma^2(\mu) = \mu$ for Poisson type data or $\sigma^2(\mu) = \mu(1-\mu)$ for binary responses. The we consider the model:

$$\begin{aligned} \eta_i &= \alpha + \int \beta(t)[X_i(t) - \mu(t)]dt, \quad i = 1, \ldots, n, \\ Y_i &= g^{-1}(\eta_i) + e_i = \mu_i + e_i \\ Ee_i &= 0, \quad E(e_i|X_i) = 0, \quad \text{var}(e_i|X_i) = \sigma^2(\mu_i). \end{aligned}$$

Here the η_i are the linear predictors, μ_i are the means and e_i are independent errors.

Since ϕ_j form a basis of the function space $L^2([0, T])$, we represent

$$\beta(t) = \sum \beta_j \phi_j(t), X(t) = \sum \xi_k \phi_k(t) \Rightarrow$$
$$\int \beta(t)[X(t) - \mu(t)]dt = \sum \beta_k \xi_k.$$

Consider sequence $M = M(n) \to \infty$ as $n \to \infty$, and M-truncated models

$$Y^{(M)} = g(\alpha + \sum_{k=1}^{M} \beta_k \xi_k) + e_i' \tilde{\sigma}(\alpha + \sum_{k=1}^{M} \beta_k \xi_k)$$

with standardized errors e_i'.

Estimating equation (Quasi-score equation)

$U(\beta) = \sum_{i=1}^n \frac{(y_i - \mu_i)}{g'(g^{(-1)}(\eta_i))\sigma^2(\mu_i)} \xi_{ij} = 0, \quad j = 1, \ldots, M$, solved for $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_M \end{pmatrix}$ by Newton-Raphson (iterated weighted least squares).

Asymptotic convergence as $M \to \infty$, $\hat{\beta}(t) = \sum_{k=1}^M \hat{\beta}_k \hat{\phi}_k(t) \to_p \beta(t)$.

Application: Classification of random functions (e.g., temporal gene expressions):

Data (X_i, C_i), class label $C_i = \begin{cases} 0 \\ 1 \end{cases}$, a binary outcome.

Fit model from a training set, usually by choosing $g \equiv$ logit (logistic functional discriminant analysis). Then apply the model to a test data set, obtain $\hat{P}(C_i = 1|X_i) = \hat{E}(Y_i|X_i)$ and classify based on these estimated probabilities.

25.3 PROBLEM SET 6

1. For the toxicity data TOX.DAT (D=Dilantin, P= PM334; dose (mg/kg); no. of subjects assigned; reactions observed) for the anti-convulsive drugs Dilantin and PM334, fit a logit model for log doses and test for log-parallelity. If log-parallelity is rejected, can you think of alternative methods to compare the dose-response curves (answer this in any case but carry it out only if log-parallelity is, indeed, rejected).

2. For the anti-convulsive effect data CONV.DAT (same structure as TOX.DAT), fit a probit model for log doses and test for log-parallelity. Establish a 95% c.i. for the relative strength q.

3. Use the delta method to prove the following two results. Give all details, starting with Taylor expansion and asymptotic normality of the original parameter estimates.

   (a) An approximate $100(1 - \gamma)$ c.i. for the $ED\alpha$ is given by

$$E\hat{D}\alpha \pm \Phi^{-1}(1-\gamma/2)\left\{\frac{v\hat{a}r(\hat{\beta}_0)}{\hat{\beta}_1^2}+2\left(\frac{F^{-1}(\alpha)-\hat{\beta}_0}{\hat{\beta}_1^3}\right)c\hat{o}v(\hat{\beta}_0,\hat{\beta}_1)+\left(\frac{F^{-1}(\alpha)-\hat{\beta}_0}{\hat{\beta}_1^2}\right)^2 v\hat{a}r(\hat{\beta}_1)\right\},$$

with further details provided in class.

(b) An approximate $100(1-\gamma)$ c.i. for the relative strength q is given by

$$[\hat{q}/\exp\{\Phi^{-1}(1-\gamma/2)\hat{v}\}, \hat{q}\exp\{\Phi^{-1}(1-\gamma/2)\hat{v}\}],$$

where

$$\hat{v} = \frac{\hat{v}_{22}}{\hat{\alpha}_3^2} - 2\frac{\hat{\alpha}_2}{\hat{\alpha}_3^2}\hat{v}_{23} + \frac{\hat{\alpha}_2^2}{\hat{\alpha}_3^4}\hat{v}_{33}.$$

4. Indicate the main steps in the derivation of the asymptotic confidence interval for the relative strength q in a dose-response comparison of two log-parallel drugs. List the auxiliary results that are needed (sketch and interpret the main steps, a detailed proof is not required).

5. For a quasi-normal model with identity link, obtain (a) quasi-likelihood and quasi-score as a function of the parameter vector β, and (b) the quasi-information. How can the quasi-information be employed in model fitting?

6. A new batch of vaccine is to be prepared, and its strength is to be compared against a standard that has been used successfully in the past. You are participating as biostatistician in this effort, and your task is to determine the relative strength of the new batch as compared to the standard. Assume that five doses of vaccine from the new batch, equidistant on the log-scale, have been administered to 10 infected subjects per dose, and the same doses of the standard to another 10 infected subjects per dose, so that a total of 100 subjects are involved in the study. Success is declared when the subjects do not become diseased, i.e., the immunization appears successful. Assume throughout that we use log-doses as predictors.

   (a) Show: If the new batch is just a dilution of the standard, then one has log-parallel assays. Provide the relationship between the dilution factor and relative strength.

   (b) Characterize the properties that a possible link function needs to possess in order to be a valid link function for a GLM that is suitable for the data from this study. Name three link functions that are commonly used in dose-response analysis and give their formulas. Why are these link functions popular?

(c) Provide the components of a GLM such that log-parallelity can be tested. Write the null hypothesis (assume a general link function) formally and and describe a test statistic for carrying out this test; also give the asymptotic distribution that will be used to carry out the test.

(d) Assume a second new batch of vaccine has also been prepared, with possibly different strength as compared to the previous batch. Note that log-parallelity again is an issue in this setting. Write a linear predictor that allows to test the null hypothesis of log-parallelity between all three assays (the two assays as described above and the assay obtained from the second new batch of vaccine). Formally write the null hypothesis of log-parallelity between all three assays and indicate how one would test it.

7. For the dilution assay model, derive two versions of confidence intervals for N, the no. of agents per volume, where the estimate $\hat{N} = \exp(\hat{\beta}_0)$. You may assume a basic asymptotic normality result

$$\sqrt{n}(\hat{\beta}_0 - \beta_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$$

8. In a dose-response experiment to study the effect of anti-depressants in psychiatry, the goal is to analyze the joint action of two drugs. The drugs are X and Y, and the doses given are d_X and d_Y. The number of combinations of doses that are included is 64, where each of these dose combinations is given to one subject. The response is whether or not at least one episode of clinical depression occurs during a 3-month observation period. The plan is to fit a GLM with linear predictor $\eta = \beta_0 + \beta_1 d_X + \beta_2 d_Y$.

(a) Write all components of the GLM with general link function.

(b) Characterize the link functions that can be used for this model (necessary and sufficient criteria). How would you choose a link function in practice?

(c) Assume it is decided to use the probit link. Write the term that is minimized for updating the parameter vector the iterated weighted least squares algorithm. Provide details for the weight matrix W, plugging in the specific quantities that correspond to the model specification.

(d) Show how the weighted least squares step can be implemented with a regular least squares program that does not include a weighted least squares option.

(e) Using fitted parameters, provide an estimate of the set of dose levels $\{d_X, d_Y\}$ that correspond to the ED50.

(f) Express the joint $(1-\alpha)$-confidence ellipsoid for parameters $(\beta_0, \beta_1, \beta_2)$ in terms of the asymptotic distributional properties of the parameter estimates. Provide also an estimate for this confidence ellipsoid. Then use this confidence ellipsoid to provide a construction of a confidence region for the set of dose levels $\{d_X, d_Y\}$ that correspond to the ED50.

(g) An investigator asks you to test for the notion that drugs X and Y reinforce each other, for an enhanced joint effect, as compared to the model considered so far. Indicate how you could extend the above model, formulate null hypothesis and alternative, (No need to specify test further).

9. In an epidemiological study, randomly sampled male subjects are entered, age (in years) and smoking behavior (average no. of cigarettes per day in previous 10 years) are ascertained. Subjects are then monitored for one year to determine whether a diagnosis of lung cancer is made during this follow-up period (yes/no). Your epidemiological collaborator would like you to assess the impact of the predictors on the outcome. You decide to use a GLM type approach and the logistic link function. The sample size is n.

   (a) Write all components of a GLM.

   (b) Set up the quasi-likelihood (QL) and the QL estimating equation.

   (c) Compare the QL estimating equation with the GLM estimating equation obtained from the exponential family assumption.

   (d) (a) Obtain an estimate for the odds ratio ψ of smoking (more precisely, the odds ratio of smoking one additional cigarette per day) and also (b) the 95% confidence interval for the odds ratio. (c) Provide a justification for the confidence interval, starting with asymptotic results (no proof required).

   (e) Your collaborator is only interested to establish that $\psi > 1$ and asks you to provide a p-value for this. State null hypothesis and alternative (a) in terms of ψ and (b) in terms of the GLM parameters. (c) State the distributional result on which the test can be based.

   (f) (a) Give at least three other link functions you could use instead of the logistic. (b) What is the advantage of using the logistic link in this setting?

   (g) Your collaborator thinks that age may alter the impact of smoking on lung cancer. (a) Develop a GLM type model that allows you to check this out and write a null hypothesis for testing this. (b) Which tests can be used?

   (h) Assume that there is evidence that the effects of smoking and age are nonlinear rather than linear. Write an alternative model that corresponds to GAM (Generalized Additive Modeling).

10. In a clinical study on end-stage renal disease, one wants to investigate the relationship of patient characteristics with disease outcomes. Specifically, data on (a) years on dialysis (Y), (b) Albumin level (ALB) and (c) Age (AGE) are obtained for a sample

of 200 patients who receive dialysis. The outcome variable is whether these patients have signs of malnutrition (yes or no). Consider to use a GLM or a GAM to model the relationship of the malnutrition outcome with the predictors.

(a) Write all components for analyzing these data for (a) GLM (b) GAM.

(b) Show that GAM is a generalization of GLM.

(c) Explain why GAM overcomes the "curse of dimension" which affects other nonparametric approaches.

(d) Provide details about the iterations used to implement GAM, when the Gauss-Seidel iteration is included, as for example in R gam.

(e) When you present the results from GAM, which items do you include and how are they interpreted?

(f) If the true relationship is $g(E(Y|X = x)) = \alpha + f_1(\mathrm{Y}) + \beta\mathrm{AGE} \times \mathrm{ALB}$, how could you proceed to obtain a reasonable fit?

(g) Assume that Y, ALB and AGE are independent random variables. We now want to fit an additive model (AM) for a continuous outcome. If the basic assumptions for the AM (which are the same as for GAM except that the link function is the identity function) are satisfied, show rigorously that in this case the estimates for the component functions f_1, f_2, f_3 can be easily obtained by three simple one-dimensional smoothing steps (i.e., nonparametric regressions with continuous response versus a one-dimensional predictor).

(h) Discuss the advantages and disadvantages of GAM as compared to GLM.

11. A microbiology lab needs to determine the concentration of staphylococcus aureus (staph) bacteria $N/volumeunit$ in a blood sample. The lab conducts a 1:1 dilution assay.

(a) (5) Justify the assumption that the number of bacterial growth patches observed at each dilution level is Poisson distributed with parameter $\alpha_k = N/2^k$.

(b) (5) Use this fact to derive the model $\log(-\log(1 - \mu_k)) = \beta_0 - k \log 2$ where $\beta_0 = \log(N)$, and where μ_k is the probability to observe bacterial growth at the k-th dilution level.

(c) (5) Describe what kind of model this is and its components.

(d) (10) One requires a 95% lower confidence bound for N. Derive a formula and provide details how to obtain the unknown quantities.

(e) (10) Assume an investigator raises the possibility that the probability of a growth being observed at the $k-$th dilution level also depends on the temperature which cannot be exactly controlled and instead is measured (T_k) at each dilution level.

Write an extended model that includes this additional predictor. How would you proceed to address the concern of the investigator?

12. In a study of the effects of a new drug on the frequency of epileptic seizures, one records duration of disease, age, number of seizures in the previous 6 months and then records the number of seizures in the subsequent 6 months for a group of patients diagnosed with a specific form of epilepsy. The patients are randomly assigned to two treatment groups, and the study is double-blind. One group receives the new drug, the other group a placebo, in addition to any previous medications that the patients continue to receive.

    (a) Write the components of an appropriate GLM for this study, in the form as you would write it for a report of a statistical analysis intended to become part of a medical publication (choose the log link function).

    (b) Assume you detect overdispersion. (a) Write the variances of the responses for two common GLMs for such data (derivation not required). (b) Briefly discuss how you would decide which model to use.

    (c) It is proposed to use a variance function of the form $V(\mu) = \mu^\gamma$ with an additional parameter $\gamma > 0$. (a) For each given γ, write the quasi-likelihood and estimating equation (no explicit calculations required). (b) For which values of λ is there a corresponding GLM? Identify these GLMs. (c) How would you determine the value of γ for the case where it is unknown? Describe a possible approach in one sentence.

13. A clinical trial comparing three drugs for the treatment of arthritis has been conducted in the form of a double-blind, controlled randomized study, for which 240 elderly female patients have been recruited. The question is whether an established drug A can be replaced by alternatives B or C, which are known to have less side effects, and what the effect of the dosage is on the outcome. The outcome is measured as a rating of movement and pain experienced by patients during a monitoring period. This rating is then converted into a 0-1 recording (failure or success) for each patient. Each patient is randomized to receive one of four log-doses for one of the three drugs, so that there are 12 treatment groups in total. Randomization is balanced in such a way that there are 20 subjects per treatment group.

    (a) Define all predictors needed for an appropriate statistical model.

    (b) Write the components of a probit model.

    (c) Develop a null hypothesis for log-parallelity of all three drugs by first defining a suitable mode for the linear predictor, in which one can test this hypothesis.

Write the null hypothesis. Which test statistic could be used to carry out this test?

(d) We are now discarding drug C from consideration and would like to estimate the relative strength q of drug B relative to drug A. Assume that the effects of A and B are log-parallel. Obtain an estimate of q.

14. In order to determine whether there is a link between fat intake in the diet and breast cancer in women, $n = 150$ female twins, where exactly one of the two twins has been diagnosed with breast cancer, were studied. In this study the fat intake in the diet (fat calories) is measured by asking the enrolled subjects to fill out a daily food questionnaire for a month, then taking the average of the quotients of daily fat calories over total calories consumed. This subject-specific average quotient is referred to as fat intake in the following. In addition to fat intake, a second covariate that has been recorded is age for each twin pair.

    (a) Write the conditional likelihood for a given matched set in the general form (do not simplify). What are the variables in the model?

    (b) If one simplifies the conditional likelihood for the case of a logistic link, writing x_{i1} for the fat intake of the twin with breast cancer and x_{i0} for the twin without cancer, and takes the likelihood across all matched sets, one finds

    $$L = \prod_{j=1}^{n} \Big[1 + \sum_{i=1}^{M} \exp\{ \sum_{k=1}^{p-1} \beta_k (x_{ijk} - x_{0jk}) \} \Big]^{-1}$$

    i. Rewrite the above general formula for the case of this study, i.e., substitute M, p and n. How is the dependency between the twins belonging to the same twin pair handled in this approach?

    ii. What is the natural null hypothesis and alternative that allow to address the major question of this research? Formulate these hypotheses both scientifically and formally in terms of suitable parameters in your model. Give details of your reasoning.

    iii. How can one construct a confidence region for the odds ratio ψ when fat intake is increased by one unit? You do not need to carry out the details of the construction but you need to write the features of the model that allow to quantify the odds ratio. In which way does the odds ratio depend on the unit in which fat intake is measured?

    (c) What are the advantages/disadvantages of matched case-control studies, in comparison with regular case-control studies? When comparing with cohort studies?