Math 533
Extra Hour Material

A Justification for Regression

The Justification for Regression

It is well-known that if we want to predict a random quantity Y using some quantity m according to a mean-squared error MSE, then the optimal predictor is the expected value of Y, μ;

$$\begin{aligned}\mathbb{E}[(Y-m)^2] &= \mathbb{E}[(Y-\mu+\mu-m)^2] \\ &= \mathbb{E}[(Y-\mu)^2] + \mathbb{E}[(\mu-m)^2] + 2\mathbb{E}[(Y-\mu)(\mu-m)] \\ &= \mathbb{E}[(Y-\mu)^2] + (\mu-m)^2 + 0\end{aligned}$$

which is minimized when $m = \mu$.

The Justification for Regression (cont.)

Now suppose we have a joint distribution between Y and X, and wish to predict the Y as a function of X, $m(X)$ say. Using the same MSE criterion, if we write

$$\mu(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

to represent the conditional expectation of Y given $X = x$, we have

$$\begin{aligned}
\mathbb{E}_{X,Y}[(Y - m(X))^2] &= \mathbb{E}_{X,Y}[(Y - \mu(X) + \mu(X) - m(X))^2] \\
&= \mathbb{E}_{X,Y}[(Y - \mu(X))^2] + \mathbb{E}_{X,Y}[(\mu(X) - m(X))^2] \\
&\quad + 2\mathbb{E}_{X,Y}[(Y - \mu(X))(\mu(X) - m(X))] \\
&= \mathbb{E}_{X,Y}[(Y - \mu(X))^2] + \mathbb{E}_{X}[(\mu(X) - m(X))^2] + 0
\end{aligned}$$

The Justification for Regression (cont.)

The cross term equates to zero by noting that by iterated expectation

$$\mathbb{E}_{X,Y}[(Y-\mu(X))(\mu(X)-m(X))] = \mathbb{E}_X[\mathbb{E}_{Y|X}[(Y-\mu(X))(\mu(X)-m(X))]]$$

and for the internal expectation

$$\mathbb{E}_{Y|X}[(Y-\mu(X))(\mu(X)-m(X))|X] = (\mu(X)-m(X))\mathbb{E}_{Y|X}[(Y-\mu(X))|X]$$

and

$$\mathbb{E}_{Y|X}[(Y-\mu(X))|X] = 0 \quad \text{a.s.}$$

The Justification for Regression (cont.)

This the MSE is minimized over functions $m(.)$ when

$$\mathbb{E}_X[(\mu(X) - m(X))^2]$$

is minimized, but this term can be made zero by setting

$$m(x) = \mu(x) = \mathbb{E}_{Y|X}[Y|X = x].$$

Thus the MSE-optimal prediction is made by using $\mu(x)$.

Note: Here X can be a single variable, or a vector; it can be random or non-random – the result holds.

The Justification for Linear Modelling

Suppose that the true conditional mean function is represented by $\mu(x)$ where x is a single predictor. We have by Taylor expansion around $x = 0$ that

$$\mu(x) = \mu(0) + \sum_{j=1}^{p-1} \frac{\mu^{(j)}(0)}{j!} x^j + O(x^p)$$

where the remainder term $O(x^p)$ represents terms of x^p in magnitude or higher order terms, and

$$\mu^{(j)}(0) = \left. \frac{d^j \mu(x)}{dx^j} \right|_{x=0}$$

The Justification for Linear Modelling (cont.)

The derivatives of $\mu(x)$ at $x = 0$ may be treated as unspecified constants, in which case a reasonable approximating model takes the form

$$\beta_0 + \sum_{j=1}^{p-1} \beta_j x^j$$

where $\beta_j \equiv \mu^{(j)}(0)$ for $j = 0, 1, \ldots, p-1$.

Similar expansions hold if x is vector valued.

The Justification for Linear Modelling (cont.)

Finally, if Y and X are jointly normally distributed, then the conditional expectation of Y given $X = x$ is linear in x.

- see Multivariate Normal Distribution handout.

The General Linear Model

The General Linear Model

The linear model formulation that assumes

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta \qquad \mathbb{V}\text{ar}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \sigma^2\mathbf{I}_n$$

is actually quite a general formulation as the rows $\mathbf{x}_i$ of $\mathbf{X}$ can be formed by using general transforms of the originally recorded predictors.

- multiple regression: $\mathbf{x}_i = [1\ x_{i1}\ x_{i2}\ \cdots\ x_{ik}]$
- polynomial regression: $\mathbf{x}_i = [1\ x_{i1}\ x_{i1}^2\ \cdots\ x_{i1}^k]$

The General Linear Model (cont.)

- harmonic regression: consider single continuous predictor x measured on a bounded interval.

  Let $\omega_j = j/n, j = 0, 1, \ldots, n/2 = K$, and then set

$$\mathbb{E}_{Y_i|X}[Y_i|x_i] = \beta_0 + \sum_{j=1}^{J} \beta_{j1} \cos(2\pi\omega_j x_i) + \sum_{j=1}^{J} \beta_{j2} \sin(2\pi\omega_j x_i)$$

  If $J = K$, we then essentially have an $n \times n$ matrix $\mathbf{X}$ specifying a linear transform of $\mathbf{y}$ in terms of the derived predictors

$$(\cos(2\pi\omega_j x_i), \sin(2\pi\omega_j x_i)).$$

  – the coefficients

$$(\beta_0, \beta_{11}, \beta_{12}, \ldots, \beta_K)$$

  form the discrete Fourier transform of $\mathbf{y}$

The General Linear Model (cont.)

In this case, the columns of $\mathbf{X}$ are orthogonal, and

$$\mathbf{X}^\top\mathbf{X} = \text{diag}(n, n/2, n/2, \ldots, n/2, n)$$

that is, $\mathbf{X}^\top\mathbf{X}$ is a **diagonal** matrix.

The General Linear Model (cont.)

- basis functions:
  - truncated spline basis: for $x \in \mathbb{R}$, let

$$x_{i1} = \begin{cases} (x-\eta_1)^\alpha & x > \eta_1 \\ 0 & x \leq \eta_1 \end{cases} = (x-\eta_1)_+^\alpha$$

for some fixed η_1, and $\alpha \in \mathbb{R}$. More generally,

$$\mathbb{E}_{Y_i|X}[Y_i|x_i] = \beta_0 + \sum_{j=1}^{J} \beta_j (x_i - \eta_j)_+^\alpha$$

for fixed $\eta_1 < \eta_2 < \cdots < \eta_J$.

The General Linear Model (cont.)

- piecewise constant: : for $x \in \mathbb{R}$, let

$$x_{i1} = \begin{cases} 1 & x \in A_1 \\ 0 & x \notin A_1 \end{cases} = \mathbb{1}_{A_1}(x)$$

for some set A_1. More generally,

$$\mathbb{E}_{Y_i|X}[Y_i|x_i] = \beta_0 + \sum_{j=1}^{J} \beta_j \mathbb{1}_{A_j}(x)$$

for sets $A_1, \ldots, A_J$. If we want to use a partition of $\mathbb{R}$, we may write this

$$\mathbb{E}_{Y_i|X}[Y_i|x_i] = \sum_{j=0}^{J} \beta_j \mathbb{1}_{A_j}(x).$$

where

$$A_0 = \left(\bigcup_{j=1}^{J} A_j \right)'$$

The General Linear Model (cont.)

- piecewise linear: specify

$$\mathbb{E}_{Y_i|X}[Y_i|x_i] = \sum_{j=0}^{J} \mathbb{1}_{A_j}(x)(\beta_{j0} + \beta_{j1}x_i)$$

- piecewise linear & continuous: specify

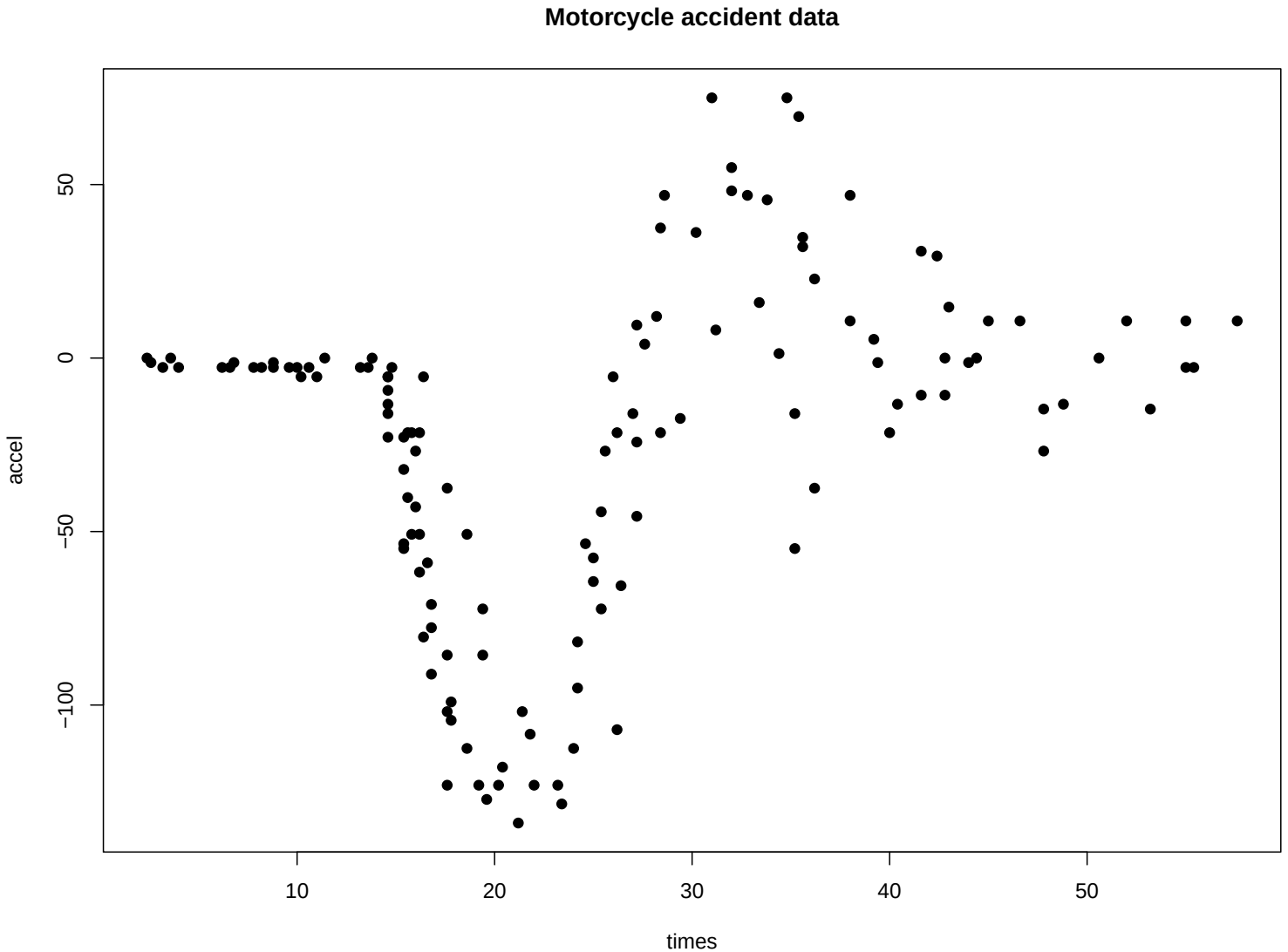
$$\mathbb{E}_{Y_i|X}[Y_i|x_i] = \beta_0 + \beta_1 x + \sum_{j=1}^{J} \beta_{j1}(x_i - \eta_j)_+$$

  for fixed $\eta_1 < \eta_2 < \cdots < \eta_J$.
- higher order piecewise functions (quadratic, cubic etc.)
- splines

Example: Motorcycle data

```
1  library(MASS)
2  #Motorcycle data
3  plot(mcycle,pch=19,main='Motorcycle accident data')
4
5  x<-mcycle$times
6  y<-mcycle$accel
```

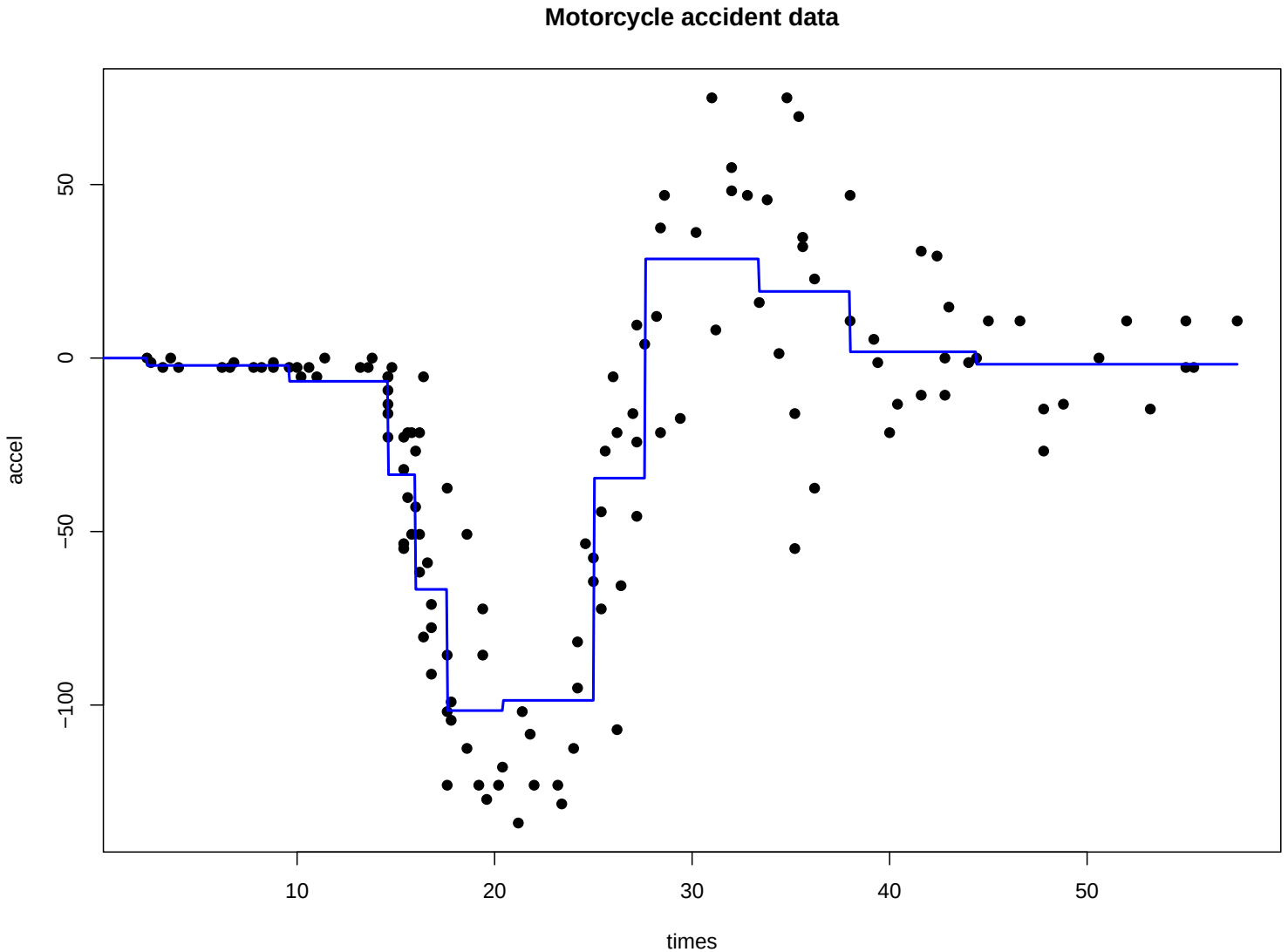
Raw data


Motorcycle accident data
accel
times
50
0
−50
−100
10
20
30
40
50

Example: Piecewise constant fit

```
#Knots
K<-11
kappa<-as.numeric(quantile(x,probs=c(0:K)/K))

X<-(outer(x,kappa,'-')>0)^2
X<-X[,-12]

fit.pwc<-lm(y~X)
summary(fit.pwc)
newx<-seq(0,max(x),length=1001)

newX<-(outer(newx,kappa,'-')>0)^2
newX<-cbind(rep(1,1001),newX[,-12])

yhatc<-newX %*% coef(fit.pwc)
lines(newx,yhatc,col='blue',lwd=2)
```

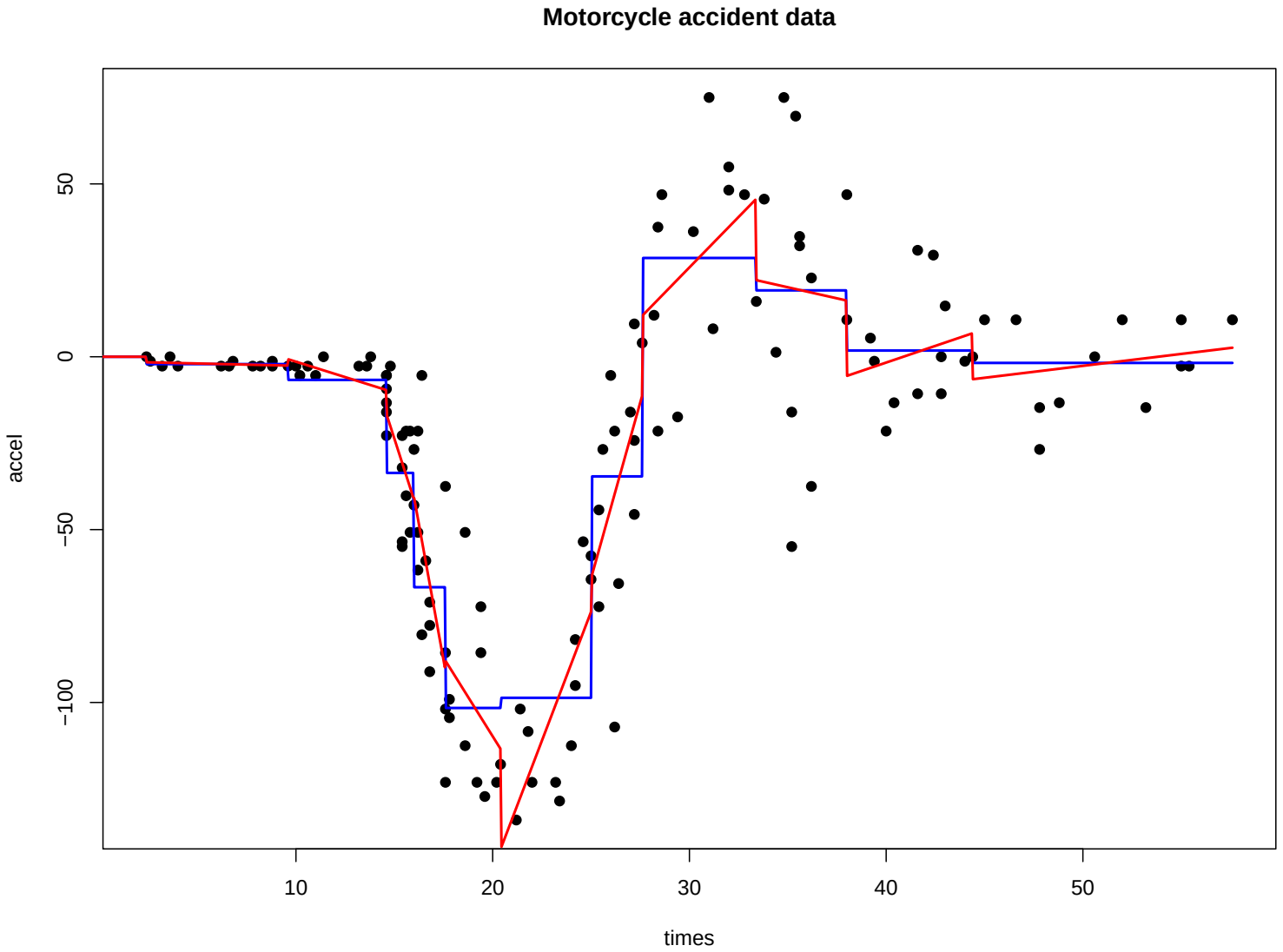
Fit: Piecewise constant



Example: Piecewise linear fit

```
17  X1<-(outer(x,kappa,'-')>0)^2
18  X1<-X1[,-12]
19
20  X2<-(outer(x,kappa,'-')>0)*outer(x,kappa,'-')
21  X2<-X2[,-12]
22
23  X<-cbind(X1,X2)
24
25  fit.pwl<-lm(y~X)
26  summary(fit.pwl)
27  newx<-seq(0,max(x),length=1001)
28
29  newX1<-(outer(newx,kappa,'-')>0)^2
30  newX2<-(outer(newx,kappa,'-')>0)*outer(newx,kappa,'-')
31
32  newX<-cbind(rep(1,1001),newX1[,-12],newX2[,-12])
33
34  yhatl<-newX %*% coef(fit.pwl)
35  lines(newx,yhatl,col='red',lwd=2)
```

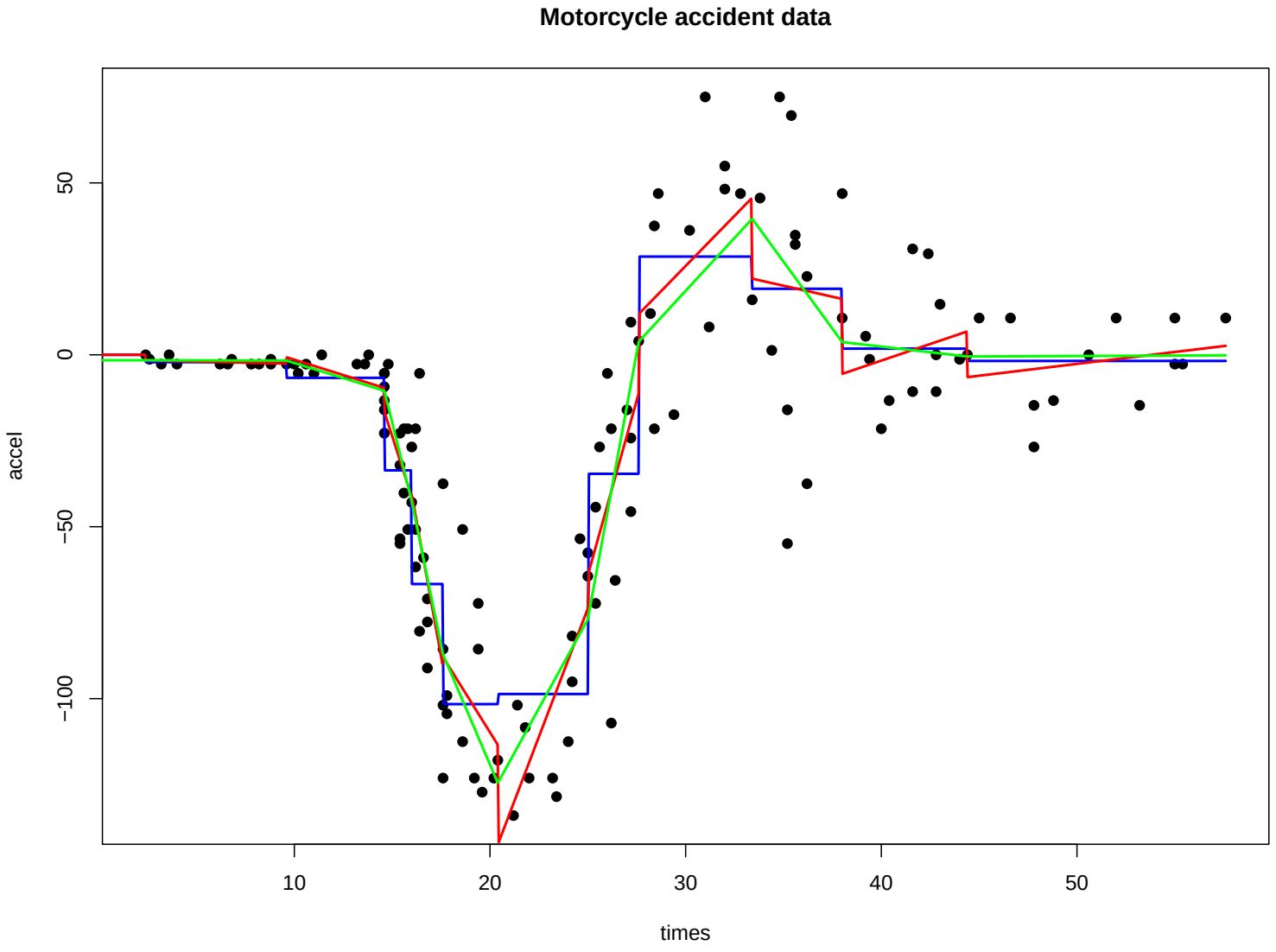
Fit: ... + piecewise linear



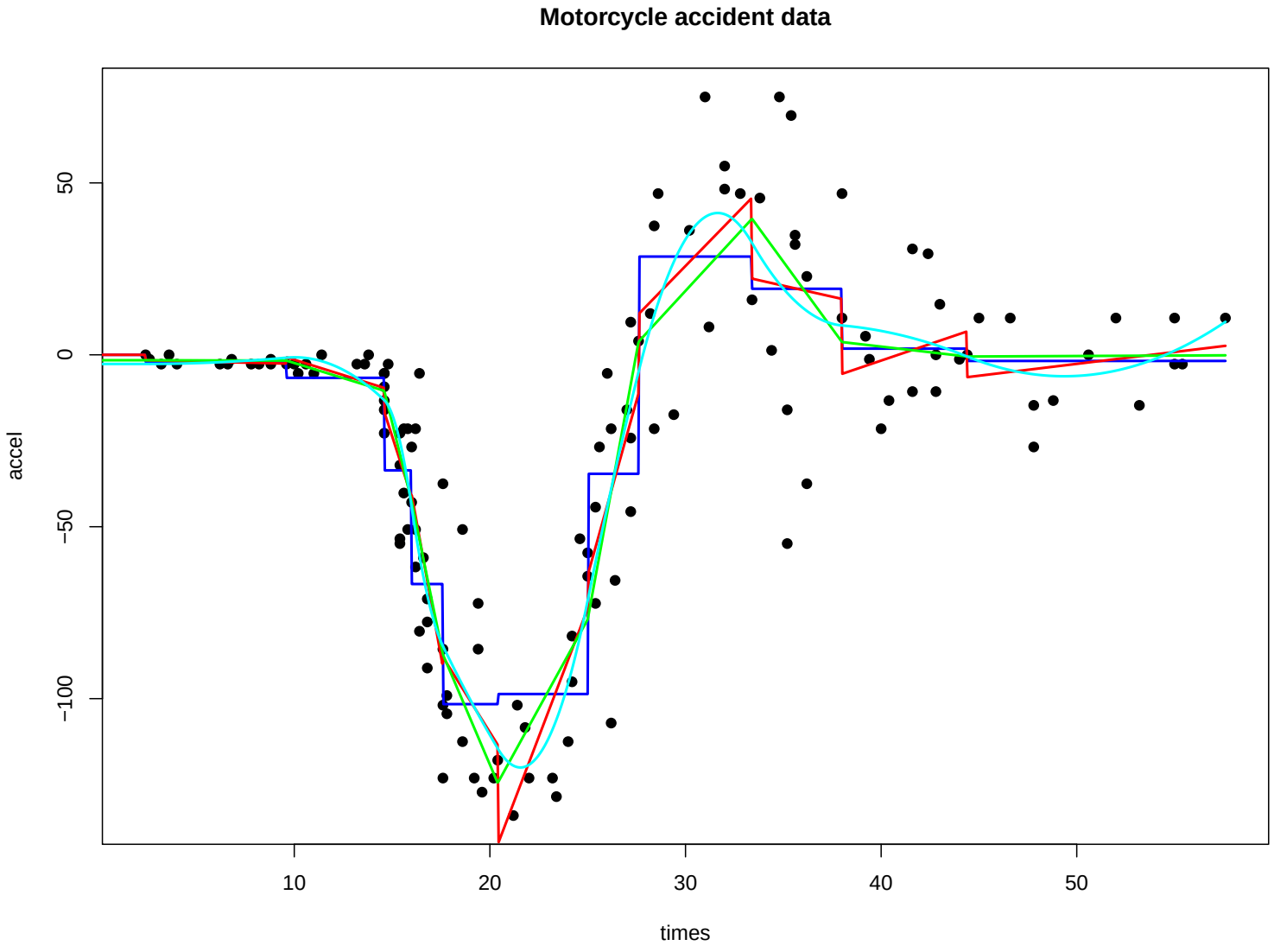
Example: Piecewise linear fit

```
36  X<-(outer(x,kappa,'-')>0)*outer(x,kappa,'-')
37  X<-X[,-12]
38
39  fit.pwcl<-lm(y~X)
40  summary(fit.pwcl)
41  newx<-seq(0,max(x),length=1001)
42
43  newX<-(outer(newx,kappa,'-')>0)*outer(newx,kappa,'-')
44
45  newX<-cbind(rep(1,1001),newX[,-12])
46
47  yhatcl<-newX %*% coef(fit.pwcl)
48  lines(newx,yhatcl,col='green',lwd=2)
```

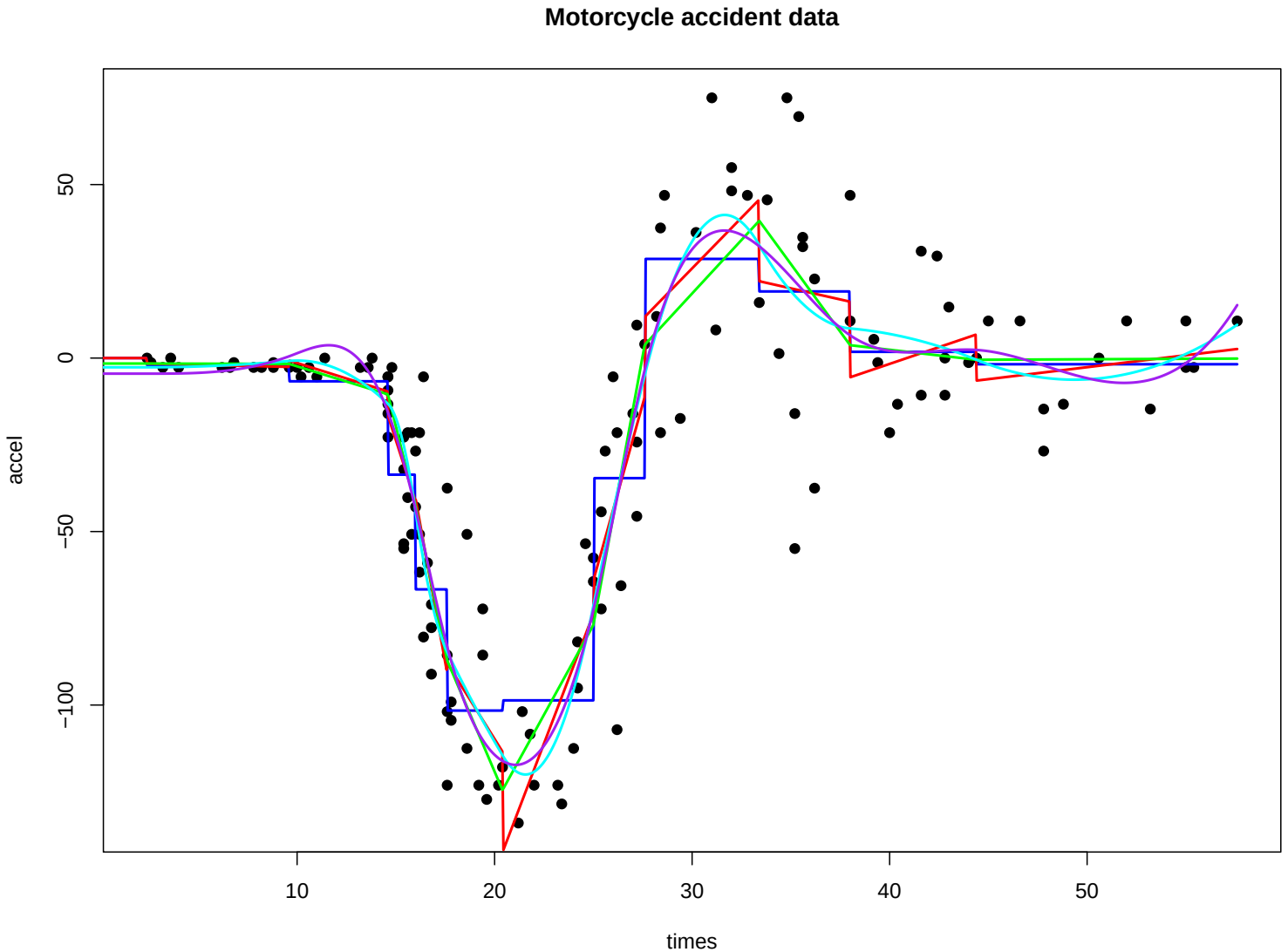
Fit: ... + piecewise continuous linear



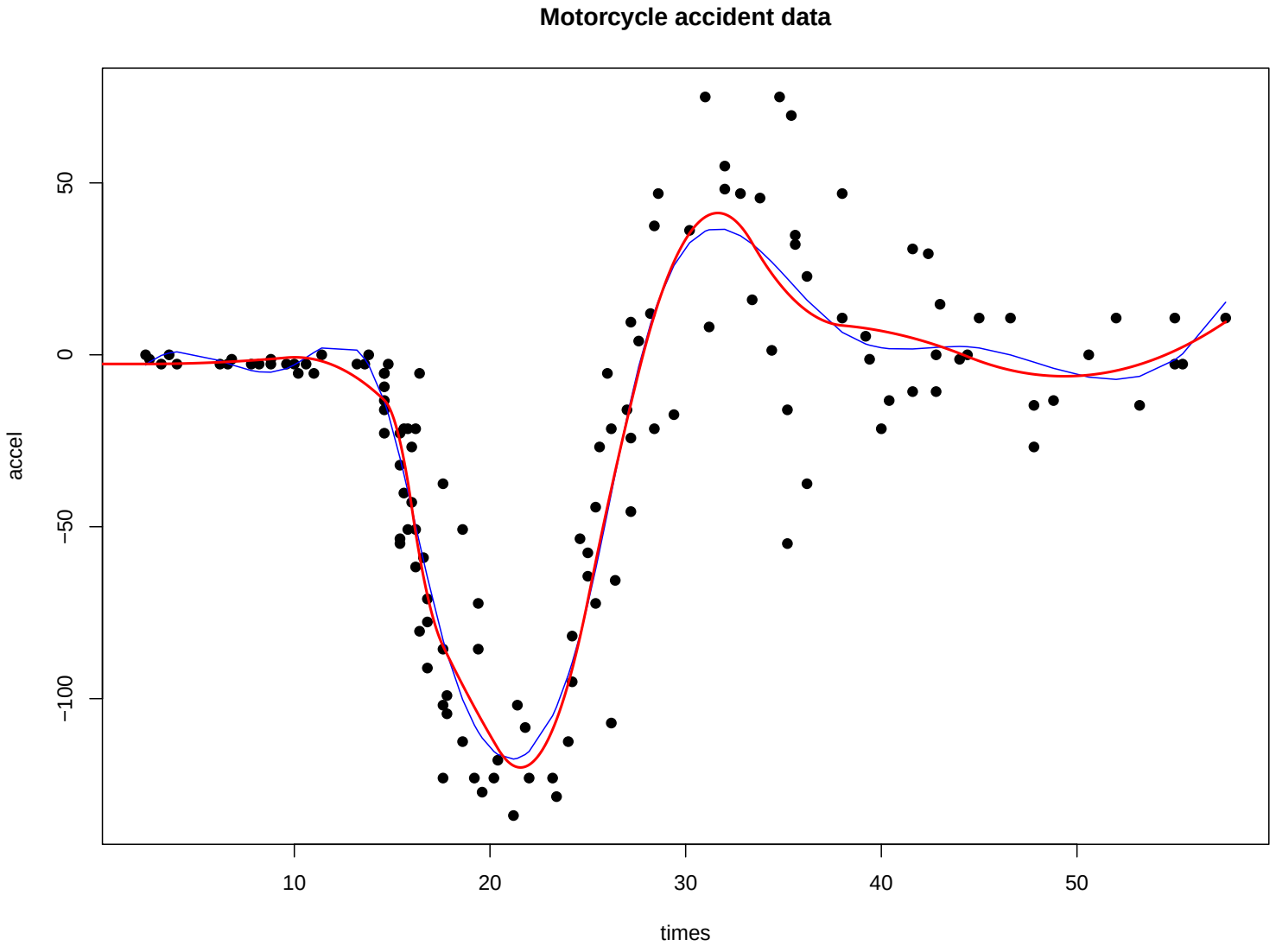
Fit: ... + piecewise continuous quadratic



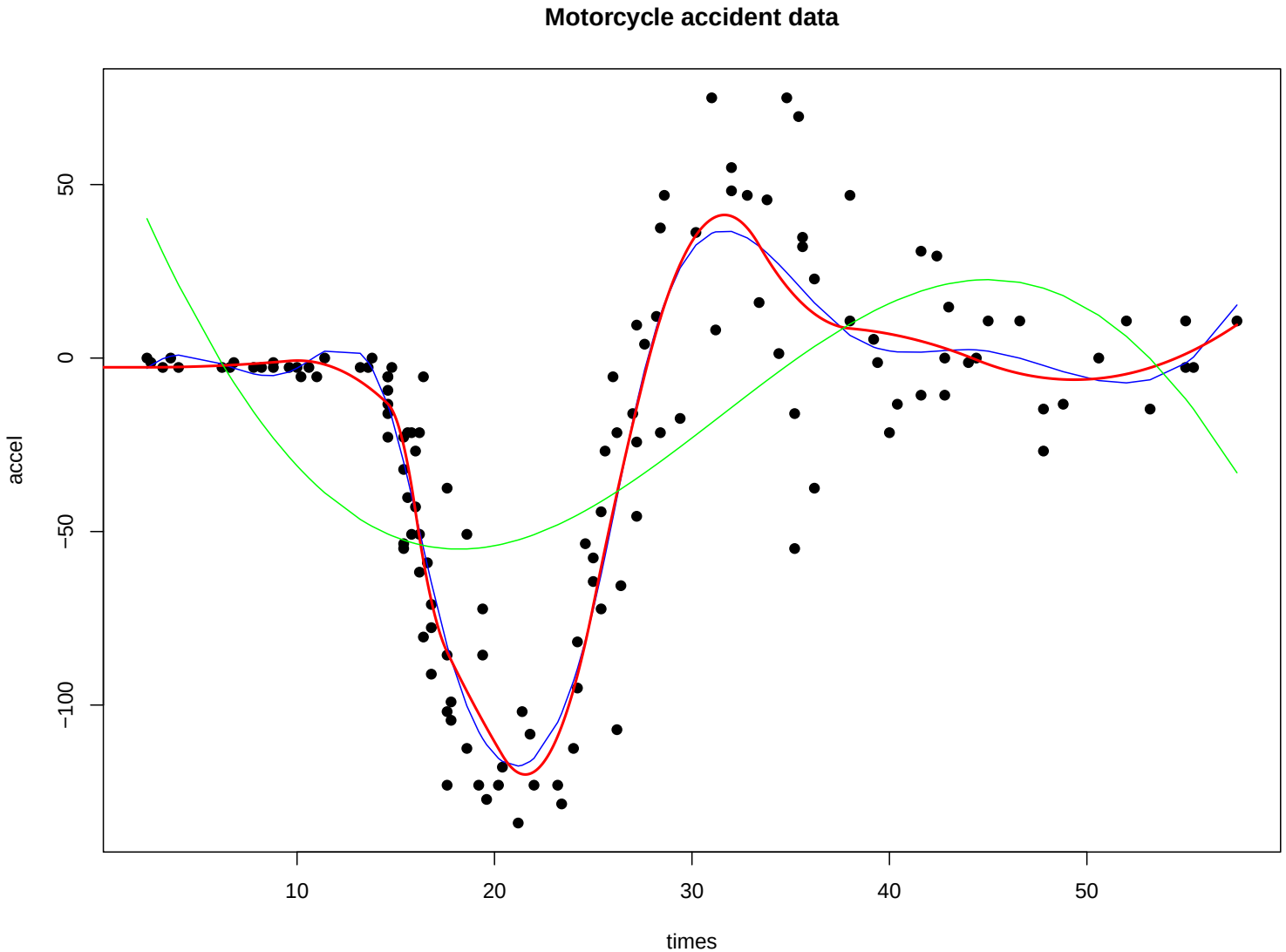
Fit: ... + piecewise continuous cubic



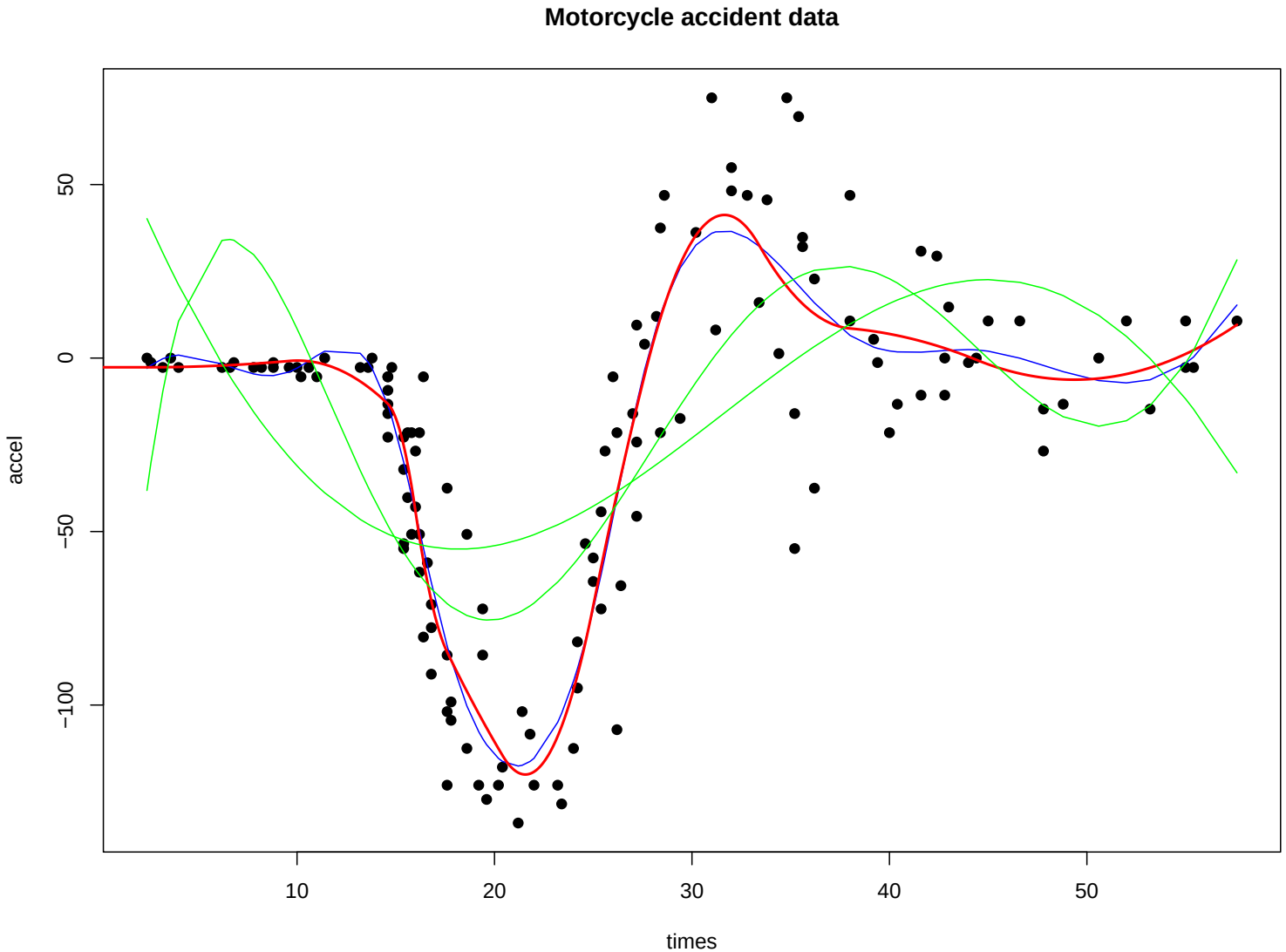
Fit: B-spline


Motorcycle accident data
accel
times
50
0
−50
−100
10
20
30
40
50

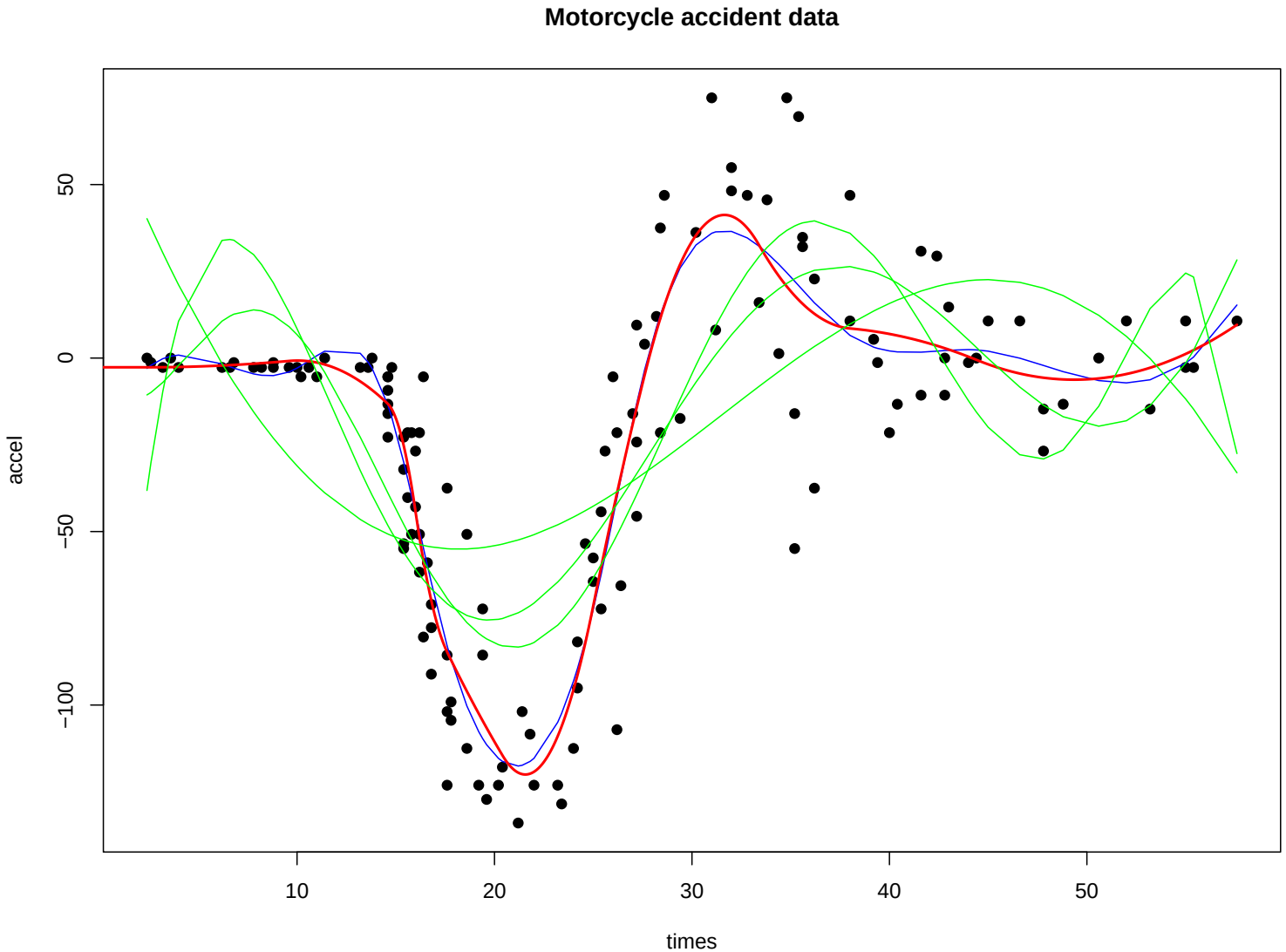
Fit: Harmonic regression $K = 2$



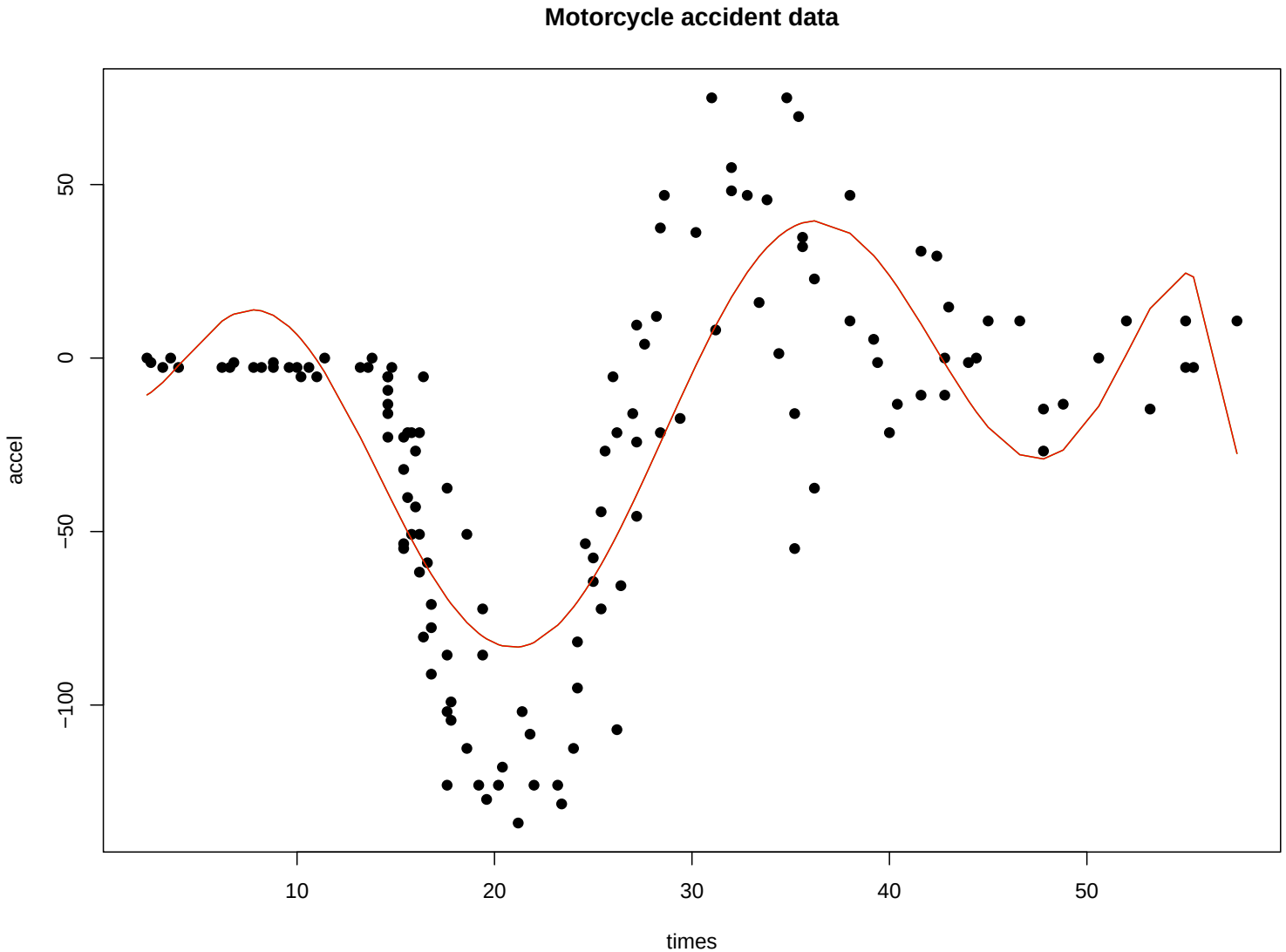
Fit: Harmonic regression $K = 3$



Fit: Harmonic regression $K = 4$



Fit: Harmonic regression $K = 5$



Reparameterization

Reparameterizing the model

For any model

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta$$

we might consider a reparameterization of the model by writing

$$\mathbf{x}_i^{\text{new}} = \mathbf{x}_i\mathbf{A}^{-1}$$

for some $p \times p$ non-singular matrix $\mathbf{A}$. Then

$$\mathbb{E}_{Y_i|X}[Y_i|\mathbf{x}_i] = \mathbf{x}_i\beta = (\mathbf{x}_i^{\text{new}}\mathbf{A})\beta = \mathbf{x}_i^{\text{new}}\beta^{(\text{new})}$$

where

$$\beta^{(\text{new})} = \mathbf{A}\beta$$

Reparameterizing the model (cont.)

Then

$$\mathbf{X}^{\text{new}} = \mathbf{X}\mathbf{A}^{-1} \qquad \Longleftrightarrow \qquad \mathbf{X} = \mathbf{X}^{\text{new}}\mathbf{A}$$

and we may choose $\mathbf{A}$ such that

$$\{\mathbf{X}^{\text{new}}\}^{\top}\{\mathbf{X}^{\text{new}}\} = \mathbf{I}_n$$

to give an orthogonal (actually, orthonormal) parameterization.

Recall that if the design matrix $\mathbf{X}^{\text{new}}$ is orthonormal, we have for the OLS estimate

$$\widehat{\beta}^{(\text{new})} = \{\mathbf{X}^{\text{new}}\}^{\top}\mathbf{y}.$$

Note however that the "new" predictors and their coefficients may not be readily interpretable, so it may be better to reparameterize back to β by defining

$$\widehat{\beta} = \mathbf{A}^{-1}\widehat{\beta}^{(\text{new})}$$

Coordinate methods for inversion

Coordinate methods for inversion

To find the ordinary least squares estimates, we solve the normal equations to obtain

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

which requires us to invert the $p \times p$ matrix

$$\mathbf{X}^\top \mathbf{X}.$$

This is the minimum norm solution to

$$\beta = \arg\min_b ||\mathbf{y} - \mathbf{X}\mathbf{b}||^2 = \arg\min_b \sum_{i=1}^{n} (y_i - \mathbf{x}_i \mathbf{b})^2$$

Coordinate methods for inversion (cont.)

We may solve this problem using *coordinate descent* rather than direct inversion; if $b_2, \ldots, b_p$ are fixed, then the minimization problem for b_1 becomes

$$\widehat{b}_1 = \arg\min_{b_1} S(b_1 | b_2, \ldots, b_p) = \arg\min_{b_1} \sum_{i=1}^{n} (y_i - b_1 x_{i1} - c_{i1})^2$$

where

$$c_{i1} = \sum_{j=2}^{p} b_j x_{ij}.$$

Coordinate methods for inversion (cont.)

Writing $y_i^* = y_i - c_{i1}$, and the sum of squares as

$$\sum_{i=1}^{n} (y_i^* - b_1 x_{i1})^2,$$

we have that

$$\widehat{b}_1 = \frac{\sum_{i=1}^{n} x_{i1} y_i^*}{\sum_{i=1}^{n} x_{i1}^2}.$$

Coordinate methods for inversion (cont.)

We may solve recursively in turn for each b_j, that is, after initialization and at step t, update

$$\widehat{b}_j^{(t)} \longrightarrow \widehat{b}_j^{(t+1)}$$

by minimizing

$$\min_{b_j} S(b_j | b_1^{(t+1)}, b_2^{(t+1)}, \ldots, b_{j-1}^{(t+1)}, b_{j+1}^{(t)} \ldots, b_p^{(t)})$$

which does not require any matrix inversion.

Distributional Results

Some distributional results

Distributional results using the Normal distribution are key to many inference procedures for the linear model. Suppose that

$$X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$$

are independent.

1. $Z_i = (X_i - \mu)/\sigma \sim \text{Normal}(0, 1)$;
2. $Y_i = Z_i^2 \sim \chi^2_1$;
3. $U = \sum_{i=1}^n Z_i^2 \sim \chi^2_n$;
4. If $U_1 \sim \chi^2_{n_1}$ and $U_2 \sim \chi^2_{n_2}$ are independent, then

$$V = \frac{U_1/n_1}{U_2/n_2} \sim \text{Fisher}(n_1, n_2)$$

and

$$\frac{1}{V} \sim \text{Fisher}(n_2, n_1)$$

Some distributional results (cont.)

5. If $Z \sim \text{Normal}(0, 1)$ and $U \sim \chi^2_\nu$, then

$$T = \frac{Z}{\sqrt{U/\nu}} \sim \text{Student}(\nu)$$

6. If

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$$

where

- $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim \text{Normal}(0, \sigma^2 \mathbf{I}_n)$;
- $\mathbf{A}$ is $n \times n$;
- $\mathbf{b}$ is $n \times 1$;

then

$$\mathbf{Y} \sim \text{Normal}(\mathbf{b}, \sigma^2 \mathbf{A}\mathbf{A}^\top) \equiv \text{Normal}(\mu, \Sigma)$$

say, and

$$(\mathbf{Y} - \mu)^\top \Sigma^{-1} (\mathbf{Y} - \mu) \sim \chi^2_n.$$

Some distributional results (cont.)

7. Non-central Chi-squared distribution: If $X \sim \text{Normal}(\mu, \sigma^2)$, we find the distribution of X^2/σ^2.

$$Y = \frac{X}{\sigma} \sim \text{Normal}(\mu/\sigma, 1)$$

By standard transformation results, if $Q = Y^2$, then

$$f_Q(y) = \frac{1}{\sqrt{y}} \left[\phi(\sqrt{y} - \mu/\sigma) + \phi(-\sqrt{y} - \mu/\sigma)\right]$$

This is the density of the non-central chi-squared distribution with 1 degree of freedom and non-centrality parameter $\lambda = (\mu/\sigma)^2$, written

$$\chi^2_1(\lambda).$$

Some distributional results (cont.)

The non-central chi-squared distribution has many similar properties to the standard (central) chi-squared distribution. For example if $X_1, \ldots, X_n$ are independent, with $X_i \sim \chi^2_1(\lambda_i)$, then

$$\sum_{i=1}^{n} X_i \sim \chi^2_n \left(\sum_{i=1}^{n} \lambda_i \right).$$

The non-central chi-squared distribution plays a role in testing for the linear regression model as it characterizes the distribution of various sums of squares terms.

Some distributional results (cont.)

8. Quadratic forms: If $\mathbf{A}$ is a square symmetric idempotent matrix, and $Z = (Z_1, \ldots, Z_n)^\top \sim \text{Normal}(0, \mathbf{I}_n)$, then

$$Z^\top \mathbf{A} Z \sim \chi^2_\nu$$

where $\nu = \text{Trace}(A)$.

To see this, use the singular value decomposition

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$

where $\mathbf{U}$ is an orthogonal matrix with $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_n$, and $\mathbf{D}$ is a diagonal matrix. Then

$$Z^\top \mathbf{A} Z = Z^\top \mathbf{U}\mathbf{D}\mathbf{U}^\top Z = (Z^\top \mathbf{U})\mathbf{D}(\mathbf{U}^\top Z) = \{Z^*\}^\top \mathbf{D}\{Z^*\}$$

say. But

$$\mathbf{U}^\top Z \sim \text{Normal}(0, \mathbf{U}^\top\mathbf{U}) = \text{Normal}(0, \mathbf{I}_n)$$

Some distributional results (cont.)

Therefore

$$\mathbf{Z}^\top \mathbf{A} \mathbf{Z} = \{\mathbf{Z}^*\}^\top \mathbf{D} \{\mathbf{Z}^*\}.$$

But $\mathbf{A}$ is idempotent, so

$$\mathbf{A}\mathbf{A}^\top = \mathbf{A}$$

that is,

$$\mathbf{U}\mathbf{D}\mathbf{U}^\top \mathbf{U}\mathbf{D}\mathbf{U}^\top = \mathbf{U}\mathbf{D}\mathbf{U}^\top.$$

The left hand side simplifies, and we have

$$\mathbf{U}\mathbf{D}^2\mathbf{U}^\top = \mathbf{U}\mathbf{D}\mathbf{U}^\top.$$

Some distributional results (cont.)

Thus, pre-multiplying by $\mathbf{U}^\top$, and post-multiplying by $\mathbf{U}$, we have

$$\mathbf{D}^2 = \mathbf{D}$$

and the diagonal elements of $\mathbf{D}$ must either be zero or one, so

$$Z^\top \mathbf{A} Z = \{Z^*\}^\top \mathbf{D} \{Z^*\} \sim \chi^2_\nu$$

where $\nu = \text{Trace}(\mathbf{D}) = \text{Trace}(\mathbf{A})$.

Some distributional results (cont.)

9. If $\mathbf{A}_1$ and $\mathbf{A}_2$ are square, symmetric and orthogonal, and

$$\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$$

then

$$Z^\top \mathbf{A}_1 Z \qquad \text{and} \qquad Z^\top \mathbf{A}_2 Z$$

are independent. This result again uses the singular value decomposition; let

$$\mathbf{V}_1 = \mathbf{D}_1\mathbf{U}_1^\top Z \qquad \mathbf{V}_2 = \mathbf{D}_2\mathbf{U}_2^\top Z.$$

We have that

$$\begin{aligned} \mathrm{Cov}_{\mathbf{V}_1,\mathbf{V}_2}[\mathbf{V}_1, \mathbf{V}_2] &= \mathbb{E}_{\mathbf{V}_1,\mathbf{V}_2}[\mathbf{V}_2\mathbf{V}_1^\top] \\ &= \mathbb{E}_Z[\mathbf{D}_2\mathbf{U}_2^\top ZZ^\top \mathbf{U}_1\mathbf{D}_1] \\ &= \mathbf{D}_2\mathbf{U}_2^\top\mathbf{U}_1^\top\mathbf{D}_1 = \mathbf{0} \end{aligned}$$

as if $\mathbf{A}_1$ and $\mathbf{A}_2$ are orthogonal, then $\mathbf{U}_2^\top\mathbf{U}_1^\top = \mathbf{0}$.

Bayesian Regression and Penalized Least Squares

The Bayesian Linear Model

Under a Normality assumption

$$\mathbf{Y}|\mathbf{X}, \beta, \sigma^2 \sim \text{Normal}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$$

which defines the likelihood $\mathcal{L}(\beta, \sigma^2; \mathbf{y}, \mathbf{X})$, we may perform Bayesian inference by specifying a joint prior distribution

$$\pi_0(\beta, \sigma^2) = \pi_0(\beta|\sigma^2)\pi_0(\sigma^2)$$

and computing the posterior distribution

$$\pi_n(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) \propto \mathcal{L}(\beta, \sigma^2; \mathbf{y}, \mathbf{X})\pi_0(\beta, \sigma^2)$$

The Bayesian Linear Model (cont.)

Prior specification:

- $\pi_0(\sigma^2) \equiv \text{InvGamma}(\alpha/2, \gamma/2)$, that is, by definition, $1/\sigma^2 \sim \text{Gamma}(\alpha/2, \gamma/2)$

$$\pi_0(\sigma^2) = \frac{(\gamma/2)^{\alpha/2)}}{\Gamma(\alpha/2)} \left(\frac{1}{\sigma^2}\right)^{\alpha/2-1} \exp\left\{-\frac{\gamma}{2\sigma^2}\right\}$$

- $\pi_0(\beta|\sigma^2) \equiv \text{Normal}(\theta, \sigma^2\Psi)$, that is, β is conditionally Normally distributed in p dimensions,

where parameters

$$\alpha, \gamma, \theta, \Psi$$

are fixed, known constants ('hyperparameters')

The Bayesian Linear Model (cont.)

In the calculation of $\pi_n(\beta, \sigma^2|\mathbf{y}, \mathbf{X})$, we have after collecting terms

$$\pi_n(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) \quad \propto \quad \left(\frac{1}{\sigma^2}\right)^{(n+\alpha+p)/2-1} \exp\left\{-\frac{\gamma}{2\sigma^2}\right\} \exp\left\{-\frac{Q(\beta, \mathbf{y}, \mathbf{X})}{2\sigma^2}\right\}$$

where

$$Q(\beta, \mathbf{y}, \mathbf{X}) = (\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) + (\beta - \theta)^\top \Psi^{-1}(\beta - \theta).$$

The Bayesian Linear Model (cont.)

By completing the square, may write

$$Q(\beta, \mathbf{y}, \mathbf{X}) = (\beta - \mathbf{m})^\top \mathbf{M}^{-1} (\beta - \mathbf{m}) + c$$

where

- $\mathbf{M} = (\mathbf{X}^\top \mathbf{X} + \Psi^{-1})^{-1}$;
- $\mathbf{m} = (\mathbf{X}^\top \mathbf{X} + \Psi^{-1})^{-1} (\mathbf{X}^\top \mathbf{y} + \Psi^{-1} \theta)$;
- $c = \mathbf{y}^\top \mathbf{y} + \theta^\top \Psi^{-1} \theta - \mathbf{m}^\top \mathbf{M}^{-1} \mathbf{m}$

The Bayesian Linear Model (cont.)

From this we deduce that

$$\pi_n(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) \propto \left(\frac{1}{\sigma^2}\right)^{(n+\alpha+p)/2-1} \exp\left\{-\frac{(\gamma + c)}{2\sigma^2}\right\} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \mathbf{m})^\top \mathbf{M}^{-1}(\beta - \mathbf{m})\right\}$$

that is

$$\pi_n(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) \equiv \pi_n(\beta|\sigma^2, \mathbf{y}, \mathbf{X})\pi_n(\sigma^2|\mathbf{y}, \mathbf{X})$$

- $\pi_n(\beta|\sigma^2, \mathbf{y}, \mathbf{X}) \equiv \text{Normal}(\mathbf{m}, \sigma^2\mathbf{M})$;
- $\pi_n(\sigma^2|\mathbf{y}, \mathbf{X}) \equiv \text{InvGamma}((n + \alpha)/2, (c + \gamma)/2)$.

The Bayesian Linear Model (cont.)

The Bayesian posterior mean/modal estimator of β based on this model is

$$\widehat{\beta}_B = \mathbf{m} = (\mathbf{X}^\top \mathbf{X} + \Psi^{-1})^{-1}(\mathbf{X}^\top \mathbf{y} + \Psi^{-1}\theta)$$

Ridge Regression

If, in the Bayesian estimation, we choose

$$\theta = 0 \qquad \Psi^{-1} = \lambda \mathbf{I}_p$$

we have

$$\widehat{\beta}_B = \mathbf{m} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}.$$

If

- $\lambda = 0$: $\widehat{\beta}_B = \widehat{\beta}$, the least squares estimate.
- $\lambda = \infty$: $\widehat{\beta}_B = 0$.

Note that to make this specification valid, we should place all the predictors (the columns of $\mathbf{X}$) on the same scale.

Ridge Regression (cont.)

Consider the constrained least squares problem

$$\text{minimize} \quad S(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i\beta)^2 \qquad \text{subject to} \quad \sum_{j=1}^{k} \beta_j^2 \leq t$$

We solve this problem after centering the predictors

$$x_{ij} \longrightarrow x_{ij} - \bar{x}_j$$

and centering the responses

$$y_i \longrightarrow y_i - \bar{y}.$$

After this transformation, the intercept can be omitted.

Ridge Regression (cont.)

Suppose that therefore there are precisely p β parameters in the model.

We solve the constrained minimization using Lagrange multipliers: we minimize $S_\lambda(\beta)$

$$S_\lambda(\beta) = S(\beta) + \lambda \left(\sum_{j=1}^{p} \beta_j^2 - t \right)$$

We have that

$$\frac{\partial S_\lambda(\beta)}{\partial \beta} = \frac{\partial S(\beta)}{\partial \beta} + 2\lambda\beta$$

– a $p \times 1$ vector.

Ridge Regression (cont.)

By direct calculus, we have

$$\frac{\partial S_\lambda(\beta)}{\partial \beta} = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta) + 2\lambda\beta$$

and equating to zero we have

$$\mathbf{X}^\top\mathbf{y} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\beta$$

so that

$$\widehat{\beta}_B = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

Ridge Regression (cont.)

For statistical properties, we have

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}_B|\mathbf{X}] = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\beta \neq \beta$$

so the ridge regression estimator is biased. However the mean squared error (MSE) of $\widehat{\beta}_{\mathrm{B}}$ can be smaller than that of $\widehat{\beta}$.

Ridge Regression (cont.)

If the columns of matrix $\mathbf{X}$ is orthogonal, so that

$$\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$$

then

$$\widehat{\beta}_{Bj} = \frac{1}{1+\lambda}\widehat{\beta}_j < \widehat{\beta}_j.$$

In general the ridge regression estimates are 'shrunk' towards zero compared to the least squares estimates.

Ridge Regression (cont.)

Using the singular value decomposition, write

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

where

- $\mathbf{U}$ is $n \times p$, columns of $\mathbf{U}$ are an orthonormal basis for the column space of $\mathbf{X}$, and

$$\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p.$$

- $\mathbf{V}$ is $p \times p$, columns of $\mathbf{V}$ are an orthonormal basis for the row space of $\mathbf{X}$;

$$\mathbf{V}^\top\mathbf{V} = \mathbf{I}_p.$$

- $\mathbf{D}$ is diagonal with elements

$$d_1 \geq d_2 \geq \cdots \geq d_p \geq 0.$$

Ridge Regression (cont.)

Least squares: Predictions are

$$\begin{aligned}\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\ &= (\mathbf{U}\mathbf{D}\mathbf{V}^\top)(\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} \\ &= \mathbf{U}\mathbf{U}^\top\mathbf{y}.\end{aligned}$$

as

$$\mathbf{V}^\top\mathbf{V} = \mathbf{I}_p \quad \Longrightarrow \quad \mathbf{V}\mathbf{V}^\top = \mathbf{I}_p$$

(to see this pre-multiply both sides by $\mathbf{V}^\top$) so that

$$(\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top)^{-1} \equiv (\mathbf{V}\mathbf{D}^2\mathbf{V}^\top)^{-1} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\top$$

Ridge Regression (cont.)

Ridge regression: Predictions are

$$\begin{aligned}\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta}_{\mathrm{B}} &= \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y} \\ &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{D}\mathbf{U}^\top\mathbf{y} \\ &= \sum_{j=1}^{p} \underset{\sim}{\mathbf{u}}_j \left(\frac{d_j^2}{d_j^2 + \lambda}\right) \underset{\sim}{\mathbf{u}}_j^\top \mathbf{y}\end{aligned}$$

where $\underset{\sim}{\mathbf{u}}_j$ is the jth column of $\mathbf{U}$.

Ridge Regression (cont.)

Ridge regression transforms the problem to one involving the orthogonal matrix $\mathbf{U}$ instead of $\mathbf{X}$, and the shrinks the coefficients by

$$\frac{d_j^2}{d_j^2 + \lambda} \leq 1.$$

For ridge regression, the hat matrix is

$$\begin{aligned} \mathbf{H}_\lambda &= \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top \\ &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{D}\mathbf{U}^\top \end{aligned}$$

and the 'degrees of freedom' of the fit is

$$\text{Trace}(\mathbf{H}_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

Other Penalties

The LASSO penalty is

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

and we solve the minimization

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j|.$$

No analytical solution is available, but the minimization can be achieved using coordinate descent.

In this case, the minimization allows for the optimal value $\widehat{\beta}_j$ to be precisely zero for some j.

Other Penalties (cont.)

A general version of this penalty is

$$\lambda \sum_{j=1}^{p} |\beta_j|^q$$

and if $q \leq 1$, there is a possibility of an estimate being shrunk exactly to zero.

For $q \geq 1$, the problem is convex; for $q < 1$ the problem is non-convex, and harder to solve.

However, if $q \leq 1$, the shrinkage to zero allows for variable selection to be carried out automatically.

Model Selection using Information Criteria

Selection using Information Criteria

Suppose we wish to choose one from a collection of models described by densities $f_1, f_2, \ldots, f_K$ with parameters $\theta_1, \theta_2, \ldots, \theta_K$. Let the true, data generating model be denoted f_0.

We consider the KL divergence between f_0 and f_k:

$$KL(f_0, f_k) = \int \log\left(\frac{f_0(x)}{f_k(x;\theta_k)}\right) f_0(x)\mathrm{d}x$$

and aim to choose the model using the criterion

$$\widehat{k} = \arg\min_k KL(f_0, f_k)$$

Selection using Information Criteria (cont.)

In reality, θ_k are typically unknown, so we consider estimating them using maximum likelihood procedures. We consider

$$KL(f_0, \widehat{f}_k) = \int \log\left(\frac{f_0(x)}{f_k(x;\widehat{\theta}_k)}\right) f_0(x)\mathrm{d}x$$

where $\widehat{\theta}_k$ is obtained by maximizing the likelihood under model k, that is maximizing

$$\sum_{i=1}^{n} \log f_k(y_i;\theta)$$

with respect to θ for data $y_1, \ldots, y_n$.

Selection using Information Criteria (cont.)

We have that

$$KL(f_0, f_k) = \int \log f_0(x) f_0(x) \mathrm{d}x - \int \log f_k(x; \theta_k) f_0(x) \mathrm{d}x$$

so we then may choose k by

$$\widehat{k} = \arg\max_k \int \log f_k(x; \theta_k) f_0(x) \mathrm{d}x$$

or if the parameters need to be estimated

$$\widehat{k} = \arg\max_k \int \log f_k(x; \widehat{\theta}_k(y)) f_0(x) \mathrm{d}x.$$

Asymptotic results for estimation under misspecification

Maximum likelihood as minimum KL

We may seek to define θ_k directly using the entropy criterion

$$\theta_k = \arg\min_\theta KL(f_0, f_k(\theta)) = \arg\min_\theta \int \log\left(\frac{f_0(x)}{f_k(x;\theta)}\right) f_0(x)\mathrm{d}x$$

and solve the problem using calculus by differentiating $KL(f_0, f_k(\theta))$ with respect to θ and equating to zero.

Note that under standard regularity conditions

$$\begin{aligned}
\frac{\partial}{\partial\theta}\left\{\int \log\left(\frac{f_0(x)}{f_k(x;\theta)}\right) f_0(x)\mathrm{d}x\right\} &= -\frac{\partial}{\partial\theta}\left\{\int \log f_k(x;\theta) f_0(x)\mathrm{d}x\right\} \\
&= -\int\left\{\frac{\partial}{\partial\theta}\log f_k(x;\theta)\right\} f_0(x)\mathrm{d}x \\
&= -\mathbb{E}_X\left[\frac{\partial}{\partial\theta}\log f_k(X;\theta)\right]
\end{aligned}$$

Maximum likelihood as minimum KL (cont.)

Therefore θ_k solves

$$\mathbb{E}_X \left[\frac{\partial}{\partial \theta} \log f_k(X;\theta)|_{\theta=\theta_k} \right] = 0.$$

Under identifiability assumptions, we assume that is that there is a single θ_k which solves this equation.

The sample-based equivalent calculation dictates that for the estimate $\widehat{\theta}_k$

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_k(x_i;\theta)|_{\theta=\theta_k} = 0$$

which coincides with ML estimation.

Maximum likelihood as minimum KL (cont.)

Under identifiability assumptions, as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f_k(X_i;\theta)|_{\theta=\theta_k} \longrightarrow 0,$$

by the strong law of large numbers we must have that

$$\widehat{\theta}_k \longrightarrow \theta_k$$

with probability 1 as $n \longrightarrow \infty$.

Maximum likelihood as minimum KL (cont.)

Let

$$\mathcal{I}(\theta_k) = \mathbb{E}_X \left[-\frac{\partial^2 \log f_k(X;\theta)}{\partial\theta\partial\theta^\top}\bigg|_{\theta=\theta_k} ; \theta_k \right]$$

be the Fisher information computed wrt $f_k(x;\theta_k)$, and

$$I(\theta_k) = \mathbb{E}_X \left[-\frac{\partial^2 \log f_k(X;\theta)}{\partial\theta\partial\theta^\top}\bigg|_{\theta=\theta_k} \right]$$

be the same expectation quantity computed wrt $f_0(x)$. The corresponding n data versions, where $\log f_k(X;\theta)$ is replaced by

$$\sum_{i=1}^{n} \log f_k(X_i;\theta)$$

are

$$\mathcal{I}_n(\theta_k) = n\mathcal{I}(\theta_k) \qquad\qquad I_n(\theta_k) = nI(\theta_k).$$

Maximum likelihood as minimum KL (cont.)

The quantity $\widehat{I}_n(\theta_k)$ is the sample based version

$$\widehat{I}_n(\theta_k) = -\left\{ \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^\top} \log f_k(X_i;\theta) \right\}_{\theta=\theta_k}$$

This quantity is evaluated at $\theta_k = \widehat{\theta}_k$ to yield $\widehat{I}_n(\widehat{\theta}_k)$; we have that

$$\widehat{I}_n(\widehat{\theta}_k) \longrightarrow I_n(\theta_k)$$

with probability 1 as $n \longrightarrow \infty$.

Maximum likelihood as minimum KL (cont.)

Another approach that is sometimes used is to uses the equivalence between expressions involving the first and second derivative versions of $\mathcal{I}(\theta)$; we have also that

$$\mathcal{I}(\theta_k) = \mathbb{E}_X \left[\frac{\partial \log f_k(X;\theta)}{\partial \theta} \Big|^{\otimes 2}_{\theta=\theta_k} ; \theta_k \right] = \mathcal{J}(\theta_k)$$

with the expectation computed wrt $f_k(x;\theta_k)$, where for vector U

$$U^{\otimes 2} = UU^\top.$$

Let

$$J(\theta_k) = \mathbb{E}_X \left[\frac{\partial \log f_k(X;\theta)}{\partial \theta} \Big|^{\otimes 2}_{\theta=\theta_k} \right]$$

be the equivalent calculation with the expectation computed wrt $f_0(x)$.

Maximum likelihood as minimum KL (cont.)

Now by a second order Taylor expansion of the first derivative of the loglikelihood, if

$$\dot{\ell}_k(\theta_k) = \left.\frac{\partial \log f_k(x;\theta)}{\partial \theta}\right|_{\theta=\theta_k} \qquad \ddot{\ell}_k(\theta_k) = \left.\frac{\partial^2 \log f_k(x;\theta)}{\partial \theta \partial \theta^\top}\right|_{\theta=\theta_k}$$

at $\theta = \theta_k$, we have

$$\begin{aligned}\dot{\ell}_k(\theta_k) &= \dot{\ell}_k(\widehat{\theta}_k) + \ddot{\ell}_k(\widehat{\theta}_k)(\theta_k - \widehat{\theta}_k) + \frac{1}{2}(\theta_k - \widehat{\theta}_k)^\top \dddot{\ell}_n(\theta^*)(\theta_k - \widehat{\theta}_k) \\ &= \ddot{\ell}_k(\widehat{\theta}_k)(\theta_k - \widehat{\theta}_k) + R_n(\theta^*)\end{aligned}$$

as $\dot{\ell}_k(\widehat{\theta}_k) = 0$, $||\widehat{\theta}_k - \theta^*|| < ||\widehat{\theta}_k - \theta_k||$, and where the remainder term is being denoted $R_n(\theta^*)$.

Maximum likelihood as minimum KL (cont.)

We have by definition that

$$-\ddot{\ell}_k(\widehat{\theta}_k) = \widehat{I}_n(\widehat{\theta}_k).$$

and in the limit

$$\frac{R_n(\theta^*)}{n} \xrightarrow{p} 0$$

that is, $R_n(\theta^*) = o_p(n)$, and

$$\frac{1}{n}\widehat{I}_n(\widehat{\theta}_k) \xrightarrow{a.s.} I(\theta_k)$$

as $n \longrightarrow \infty$.

Rewriting the above approximation, we have

$$\sqrt{n}(\widehat{\theta}_k - \theta_k) = \left\{\frac{1}{n}\widehat{I}_n(\widehat{\theta}_k)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\dot{\ell}_k(\theta_k)\right\} + \left\{\frac{1}{n}\widehat{I}_n(\widehat{\theta}_k)\right\}^{-1}\left\{\frac{R_n(\theta^*)}{n}\right\}$$

Maximum likelihood as minimum KL (cont.)

By the central limit theorem

$$\sqrt{n}(\widehat{\theta}_k - \theta_k) \xrightarrow{d} \text{Normal}(0, \Sigma)$$

say, where Σ denotes the limiting variance-covariance matrix when sampling is under f_0, and

$$\frac{1}{\sqrt{n}}\dot{\ell}_k(\theta_k) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_k}\log f_k(X_i;\theta_k) \xrightarrow{d} \text{Normal}(0, J(\theta_k)).$$

Finally,

$$\left\{\frac{1}{n}\widehat{I}_n(\widehat{\theta}_k)\right\}^{-1}\left\{\frac{R_n(\theta^*)}{n}\right\} \xrightarrow{p} 0$$

by Slutsky's Theorem.

Maximum likelihood as minimum KL (cont.)

Therefore, by equating the asymptotic variances of the above quantities, we must have

$$J(\theta_k) = I(\theta_k)\Sigma I(\theta_k)$$

yielding that

$$\Sigma = \{I(\theta_k)\}^{-1} J(\theta_k) \{I(\theta_k)\}^{-1}$$

Model Selection using Minimum KL

The Expected KL Divergence

The quantity

$$\int \log f_k(x; \widehat{\theta}_k(Y)) f_0(x) \mathrm{d}x = \mathbb{E}_X[\log f_k(X; \widehat{\theta}_k(Y))]$$

is a random quantity, a function of data random quantities Y. Thus we instead decide to choose k by

$$\widehat{k} = \arg\max_k \mathbb{E}_Y[\mathbb{E}_X[\log f_k(X; \widehat{\theta}_k(Y))]]$$

where X and Y are drawn from the true model f_0.

The Expected KL Divergence (cont.)

We consider first the expansion of the inner integral around the true value θ_k for an arbitrary y; under regularity conditions

$$\begin{aligned}\mathbb{E}_X[\log f_k(X;\widehat{\theta}_k(y))] = \; & \mathbb{E}_X[\log f_k(X;\theta_k)] \\ & + \mathbb{E}_X\left[\dot{\ell}_k(\theta_k)\right]^\top (\widehat{\theta}_k - \theta_k) \\ & + \frac{1}{2}(\widehat{\theta}_k - \theta_k)^\top \mathbb{E}_X\left[\ddot{\ell}_k(\theta_k)\right](\widehat{\theta}_k - \theta_k) \\ & + o_p(n)\end{aligned}$$

The Expected KL Divergence (cont.)

By definition $\mathbb{E}_X\left[\dot{\ell}_k(\theta_k)\right] = 0$, and

$$\mathbb{E}_X\left[\ddot{\ell}_k(\theta_k)\right] = -I_n(\theta_k)$$

so therefore

$$\mathbb{E}_X[\log f_k(X;\widehat{\theta}_k(y))] = \mathbb{E}_X[\log f_k(X;\theta_k)] - \frac{1}{2}(\widehat{\theta}_k - \theta_k)^\top I_n(\theta_k)(\widehat{\theta}_k - \theta_k) + o_p(n)$$

We then must compute

$$\mathbb{E}_Y[\mathbb{E}_X[\log f_k(X;\widehat{\theta}_k(Y))]]$$

The term $\mathbb{E}_X[\log f_k(X;\theta_k)]$ is a constant wrt this expectation.

The Expected KL Divergence (cont.)

The expectation of the quadratic term can be computed by standard results for large n; under sampling Y from f_0, we have as before that

$$(\widehat{\theta}_k - \theta_k)^\top I_n(\theta_k)(\widehat{\theta}_k - \theta_k)$$

can be rewritten

$$\{\sqrt{n}(\widehat{\theta}_k - \theta_k)\}^\top \left\{\frac{1}{n} I_n(\theta_k)\right\} \{\sqrt{n}(\widehat{\theta}_k - \theta_k)\}$$

where

$$\sqrt{n}(\widehat{\theta}_k - \theta_k) \xrightarrow{d} \text{Normal}(0, \Sigma)$$

and

$$\frac{1}{n} I_n(\theta_k) \xrightarrow{a.s.} I(\theta_k)$$

The Expected KL Divergence (cont.)

Therefore, by standard results for quadratic forms

$$\mathbb{E}_Y\left[(\widehat{\theta}_k(Y)-\theta_k)^\top I_n(\theta_k)(\widehat{\theta}_k(Y)-\theta_k)\right] \xrightarrow{a.s.} \text{Trace}\left(I(\theta_k)\Sigma\right)$$

and

$$\mathbb{E}_Y[\mathbb{E}_X[\log f_k(X;\widehat{\theta}_k(Y))]] = \mathbb{E}_X[\log f_k(X;\theta_k)] - \frac{1}{2}\text{Trace}\left(I(\theta_k)\Sigma\right) + o_p(n) \quad (*)$$

However, the right hand side cannot be computed, as θ_k is not known and must be estimated.

The Expected KL Divergence (cont.)

Thus we repeat the same operation, but instead expand

$$\mathbb{E}_X[\log f_k(X;\theta_k)]$$

about $\widehat{\theta}_k(x)$ for a fixed x. We have

$$\begin{aligned}\log f_k(x;\theta_k) = \; & \log f_k(x;\widehat{\theta}_k(x)) \\ & + \dot{\ell}_k(\widehat{\theta}_k(x))^\top(\theta_k - \widehat{\theta}_k(x)) \\ & + \frac{1}{2}(\widehat{\theta}_k(x) - \theta_k)^\top \ddot{\ell}_k(\widehat{\theta}_k(x))(\widehat{\theta}_k(x) - \theta_k) \\ & + \mathrm{o}(n)\end{aligned}$$

of which we need then to take the expectation wrt X.

The Expected KL Divergence (cont.)

Again $\mathbb{E}_X\left[\dot{\ell}_k(\widehat{\theta}_k(X))\right] = 0$, and writing

$$-(\widehat{\theta}_k(x) - \theta_k)^\top \ddot{\ell}_k(\widehat{\theta}_k(x))(\widehat{\theta}_k(x) - \theta_k)$$

as

$$\{\sqrt{n}(\widehat{\theta}_k(x) - \theta_k)\}^\top \left\{-\frac{1}{n}\ddot{\ell}_k(\widehat{\theta}_k(x))\right\}\{\sqrt{n}(\widehat{\theta}_k(x) - \theta_k)\}$$

we have that the expectation over X of this quadratic form converges to

$$\text{Trace}\,(I(\theta_k)\Sigma)$$

as by standard theory

$$\sqrt{n}(\widehat{\theta}_k(x) - \theta_k) \xrightarrow{d} \text{Normal}(0, \Sigma).$$

The Expected KL Divergence (cont.)

Thus

$$\mathbb{E}_X[\log f_k(X;\theta_k)] = \mathbb{E}_X[\log f_k(X;\widehat{\theta}_k(X))] - \frac{1}{2}\text{Trace}\left(I(\theta_k)\Sigma\right) + o_p(n).$$

Therefore, using the previous expression (*) we now have

$$\mathbb{E}_Y[\mathbb{E}_X[\log f_k(X;\widehat{\theta}_k)]] = \mathbb{E}_X[\log f_k(X;\widehat{\theta}_k(X))] - \text{Trace}\left(I(\theta_k)\Sigma\right) + o_p(n)$$

By the earlier identity

$$I(\theta_k)\Sigma = J(\theta_k)\left\{I(\theta_k)\right\}^{-1}$$

and the right hand side can be consistently estimated by

$$\log f_k(x;\widehat{\theta}_k) - \widehat{\text{Trace}}\left(J(\theta_k)\left\{I(\theta_k)\right\}^{-1}\right)$$

where the estimated trace must be computed from the available data.

The Expected KL Divergence (cont.)

On obvious estimator of the trace is

$$\text{Trace}\left(\widehat{J}(\widehat{\theta}_k(x))\left\{\widehat{I}(\widehat{\theta}_k(x))\right\}^{-1}\right)$$

although this might potentially be improved upon. In any case, if the approximating model f_k is close in KL terms to the true model f_0, we would expect that

$$\text{Trace}\left(J(\theta_k)\left\{I(\theta_k)\right\}^{-1}\right) \doteq \dim(\theta_k)$$

as we would have under regularity assumptions

$$J(\theta_k) \doteq I(\theta_k)$$

The Expected KL Divergence (cont.)

This yields the criterion: choose k to maximize

$$\log f_k(x; \widehat{\theta}_k) - \dim(\theta_k)$$

or equivalently to minimize

$$-2 \log f_k(x; \widehat{\theta}_k) + 2\dim(\theta_k)$$

This is **Akaike's Information Criterion** (AIC).

Note: The required regularity conditions on the f_k model are not too restrictive. However, the approximation

$$\text{Trace}\left(J(\theta_k)\left\{I(\theta_k)\right\}^{-1}\right) \doteq \dim(\theta_k)$$

is potentially poor.

The Bayesian Information Criterion

The Bayesian Information Criterion (BIC) uses an approximation to the marginal likelihood function to justify model selection. For data $\mathbf{y} = (y_1, \ldots, y_n)$, we have the posterior distribution for model k as

$$\pi_n(\theta_k; \mathbf{y}) = \frac{L_k(\theta_k; \mathbf{y})\pi_0(\theta_k)}{\int L_k(t; \mathbf{y})\pi_0(t)\, \mathrm{d}t}.$$

where $L_k(\theta; \mathbf{y})$ is the likelihood function. The denominator is

$$f_k(\mathbf{y}) = \int L_k(t; \mathbf{y})\pi_0(t)\, \mathrm{d}t$$

that is, the *marginal likelihood*.

The Bayesian Information Criterion (cont.)

Let

$$\ell_k(\theta; \mathbf{y}) = \log L_k(\theta; \mathbf{y}) = \sum_{i=1}^{n} \log f_k(y_i; \theta)$$

denote the log-likelihood for model k. By a Taylor expansion of $\ell_k(\theta; \mathbf{y})$ around ML estimate $\widehat{\theta}_k$, we have

$$\ell_k(\theta; \mathbf{y}) = \ell_k(\widehat{\theta}_k; \mathbf{y}) + (\theta - \widehat{\theta}_k)^\top \dot{\ell}(\widehat{\theta}_k; \mathbf{y}) + \frac{1}{2}(\theta - \widehat{\theta}_k)^\top \ddot{\ell}(\widehat{\theta}_k; \mathbf{y})(\theta - \widehat{\theta}_k) + \mathrm{o}(1)$$

We have by definition that $\dot{\ell}(\widehat{\theta}_k) = 0$, and we may write

$$-\ddot{\ell}(\widehat{\theta}_k; \mathbf{y}) = \left\{ V_n(\widehat{\theta}_k) \right\}^{-1}.$$

$V_n(\widehat{\theta}_k)$ is the Hessian matrix derived from n data points.

The Bayesian Information Criterion (cont.)

In ML theory, $V_n(\widehat{\theta}_k)$ estimates the variance of the ML estimator derived from n data. We may denote that, for a random sample, that

$$V_n(\widehat{\theta}_k) = \frac{1}{n} V_1(\widehat{\theta}_k)$$

say, where

$$V_1(\theta) = n \left\{ \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^\top} \log f_k(y_i;\theta) \right\}^{-1}$$

is a square symmetric matrix of dimension $p_k = \dim(\theta_k)$ recording an estimate of the the variance of the ML estimator for a single data point $n = 1$.

The Bayesian Information Criterion (cont.)

Then, exponentiating, we have that

$$L_k(\theta; \mathbf{y}) \doteq L_k(\widehat{\theta}_k; \mathbf{y}) \exp\left\{ -\frac{1}{2}(\theta - \widehat{\theta}_k)^\top \left\{ V_n(\widehat{\theta}_k) \right\}^{-1} (\theta - \widehat{\theta}_k) \right\}$$

This is a standard quadratic approximation to the likelihood around the ML estimate.

The Bayesian Information Criterion (cont.)

Suppose that prior $\pi_0(\theta_k)$ is constant (equal to c, say) in the neighbourhood of $\widehat{\theta}_k$. Then, for large n

$$
\begin{aligned}
f_k(\mathbf{y}) &= \int L_k(t;\mathbf{y})\pi_0(t)\,\mathrm{d}t \\
&\simeq \int cL_k(\widehat{\theta};\mathbf{y}) \exp\left\{-\frac{1}{2}(t-\widehat{\theta}_k)^\top \left\{V_n(\widehat{\theta}_k)\right\}^{-1}(t-\widehat{\theta}_k)\right\}\,\mathrm{d}t \\
&= cL_k(\widehat{\theta};\mathbf{y})(2\pi)^{p_k/2}|V_n(\widehat{\theta}_k)|^{1/2}
\end{aligned}
$$

as the integrand is proportional to a Normal$(\widehat{\theta}_k, V_n(\widehat{\theta}_k))$ distribution.

The Bayesian Information Criterion (cont.)

But

$$|V_n(\widehat{\theta}_k)|^{1/2} = \left|\frac{1}{n} V_1(\widehat{\theta}_k)\right|^{1/2} = n^{-p_k/2} \left|V_1(\widehat{\theta}_k)\right|^{1/2}$$

Therefore the marginal likelihood becomes

$$f_k(\mathbf{y}) = cL_k(\widehat{\theta}; \mathbf{y})(2\pi)^{p_k/2} n^{-p_k/2} \left|V_1(\widehat{\theta}_k)\right|^{1/2}$$

or on the $-2\log$ scale we have

$$-2\log f_k(\mathbf{y}) \doteq -2\ell_k(\widehat{\theta}; \mathbf{y}) + p_k \log n + \text{constant}$$

where

$$\text{constant} = -p_k \log(2\pi) - \log|V_1(\widehat{\theta}_k)| - 2\log c.$$

The Bayesian Information Criterion (cont.)

The constant term is $o(1)$ and is hence negligible, so the BIC is defined as

$$\text{BIC} = -2\ell_k(\widehat{\theta}; \mathbf{y}) + p_k \log n.$$