APTS course: 2nd Sept – 6th Sept 2019

Flexible Regression

Preliminary Material

Claire Miller & Tereza Neocleous

University of Glasgow | School of Mathematics & Statistics

The term 'flexible regression' refers to a wide range of methods which provide flexibility in the nature of the relationship being modelled. This APTS course will start with univariate smoothing and progress through standard forms of nonparametric regression to state-of-the art modelling tools (including quantile regression) which can be applied in a wide variety of settings.

As preparation, it would be helpful to revise the following topics covered in earlier APTS courses:

- (Bayesian) linear models (see APTS course - **Statistical Modelling**);
- generalised linear models (see APTS course - **Statistical Modelling**);
- `R` programming (see APTS course - **Statistical Computing**);
- matrix computations (see APTS course - **Statistical Modelling**);
- confidence intervals/hypothesis testing (see APTS course - **Statistical Inference**).

The main emphasis will be on regression settings, because of the widespread use and application of this kind of data structure. However, the preliminary material also covers various aspects of density estimation, to introduce some of the main ideas of nonparametric smoothing and to highlight some of the main issues involved. It is likely that many people will have come across these ideas in one form or another. The preliminary material aims to:

- revise key concepts in (generalised) linear (mixed) models and introduce notation required for the course;
- introduce the idea of smoothing through exploring simple kernel methods to construct smooth density estimates;
- revise/introduce ideas of basis functions;
- introduce the idea of regression for quantiles;
- investigate some simple theoretical properties;
- experiment with software available in `R`.

A small number of exercises are provided to assist in engaging with the material.

1 Models of interest

1.1 Flexibility in the mean

In general, for a single explanatory variable with data $x_1, \ldots, x_n$, and response data $y_1, \ldots, y_n$ we can write a regression model as:

$$Y_i = f(x_i, \boldsymbol{\beta}) + \varepsilon_i.$$

The function $f(\mathbf{x}, \boldsymbol{\beta})$ describes the relationship between the response and the predictor variable, this might take the form of a straight line or some other function, which has parameters $\boldsymbol{\beta}$. The problem is to estimate this function f. Initially, in this course we will generally assume that,

$$\mathbb{E}(\varepsilon_i) = 0 \quad \text{and} \quad \text{Var}(\varepsilon_i) = \sigma^2$$

for all i, where σ^2 does not depend on any other unknown or on x_i, and x_i are assumed to be recorded without error.

We also initially generally assume that $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$ and usually that ε_i and ε_j are uncorrelated for $i \neq j$, (i.e. independent identically distributed, i.i.d).

For Gaussian data we have a least squares loss function which we use to estimate the parameters $\boldsymbol{\beta}$ in our model, i.e. we choose $\boldsymbol{\beta}$ to minimise

$$\sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2.$$

Standard regression is sometimes referred to as mean regression, because the mean minimises the squared loss.

Previous APTS courses have considered linear and non-linear functions and the inclusion of both fixed and random effects in a regression model. In this course we will extend this by allowing $f()$ to be a data driven smooth function.

For a general smooth function $f()$ we refer to the approach as **nonparametric regression**. This extends to **(generalised) additive models** (GAMs) for more than one smooth covariate, and such models can include univariate, bivariate (or possibly higher order) terms and be extended to distributions other than the normal.

1.2 Flexibility in the response quantile

In some circumstances regression methods based on standard distributional assumptions will not capture all aspects of the distribution of the response variable of interest.

Usually regression models are based on a covariate-based model assumption for the mean only. However in some situations not just the mean, but also the spread and the shape of the distribution of the response depend on covariates. Therefore, additionally in this course we will consider **quantile regression** and combine this with approaches which allow smooth functions for the covariates, to introduce **generalised additive quantile regression models**.

When the quantity of interest is just one quantile it is easiest to fit a **quantile regression** model. Suppose we have data $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ and a predictor function $f(\mathbf{x})$ which depends on parameters $\boldsymbol{\beta}$. Instead of using the least squares loss function above, if we were to use the absolute loss

$$\sum_{i=1}^{n} |y_i - f(\mathbf{x}_i)|$$

we would obtain median regression (also known as least absolute deviations regression).

Quantile regression is based on minimising,

$$\sum_{i=1}^{n} \rho_\tau(y_i - f(\mathbf{x}_i))$$

and results in an estimate of the τ-th quantile of the response distribution, where $\rho_\tau(\cdot)$ is the so-called check function,

$$\rho_\tau(z) = \begin{cases} \tau z & \text{if z } > 0 \\ (\tau - 1)z & \text{if z} \leq 0. \end{cases}$$

This preliminary material will provide very brief revision of (generalised) linear (mixed) model concepts and introduce ideas of smoothing and quantile regression to motivate ideas for the course.

2 Revision

2.1 Linear models

In general, for explanatory variables $\mathbf{x_1}, \ldots, \mathbf{x_{p-1}}$, with response data $y_1, \ldots, y_n$ we can write our linear model as:

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

and we wish to estimate and make inferences about the parameter vector $\boldsymbol{\beta}$.

There are two main methods of parameter estimation for linear models: the methods of **least squares** and **maximum likelihood**.

Least squares

The method of least squares minimises the sum of the squares of the vertical differences between the observed values of $\mathbf{y}$ and the fitted values $\hat{\mathbf{y}}$ in order to estimate the parameter vector $\boldsymbol{\beta}$.

Minimise:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mathbb{E}(Y_i))^2$$

$$\text{i.e.} \quad S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \boldsymbol{\beta}))^2$$

with respect to $\boldsymbol{\beta}$.

This can be re-formulated as the vector-matrix form of the linear model:

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$

or

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- $\mathbf{y}$ is the $(n \times 1)$ vector of observations
- $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of parameters

- $\mathbf{X}$ is an $(n \times p)$ design matrix
- $\boldsymbol{\varepsilon}$ is the $(n \times 1)$ vector of random errors, independent and identically distributed (i.i.d) $\mathrm{N}(0, \sigma^2\mathbf{I_n})$.

Therefore, the least squares formulation can be written in vector-matrix form as:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbb{E}(Y_i))^2 = (\mathbf{y} - \mathbb{E}(\mathbf{y}))^\top(\mathbf{y} - \mathbb{E}(\mathbf{y})) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Differentiating with respect to $\boldsymbol{\beta}$ and setting to zero, we get the normal equations:

$$\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{y}.$$

Provided $\mathbf{X}$ is an invertible matrix (i.e. positive definite), this gives:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

From this, we can derive straight forward expressions for fitted values, residuals and the residual sum of squares.

The **fitted values** are derived as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{H}\mathbf{y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

$\mathbf{H}$ is often referred to as the 'hat' or projection matrix, where $\mathbf{H}^\top\mathbf{H} = \mathbf{H}^2 = \mathbf{H}$. The diagonal elements of $\mathbf{H}$, H_{ii} are called leverages and are useful diagnostics.

The **residuals** (our estimate of the error term) are, therefore,

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

and the **residual sum-of-squares** (RSS) can be written as:

$$\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \mathbf{H}\mathbf{y})^\top(\mathbf{y} - \mathbf{H}\mathbf{y}) = ||\mathbf{y} - \mathbf{H}\mathbf{y}||^2 = \mathbf{y}^\top(\mathbf{I} - \mathbf{H})\mathbf{y}$$

where $||\mathbf{v}|| = \sqrt{\mathbf{v}^\top\mathbf{v}}$.

Typically the standard deviations of residuals in a sample vary greatly from one data point to another even when the errors all have the same standard deviation. We can adjust for this problem by using modified residuals i.e. standardised residuals or studentised residuals.

Standardised residuals:

$$r_i = \frac{\hat{\varepsilon_i}}{\sqrt{\text{var}(\hat{\varepsilon_i})}}.$$

For this course it will also be particularly useful for us to examine the partial effects of individual covariates. Therefore, we define,

Partial residuals, $\hat{\boldsymbol{\varepsilon}}_p$:

$$\hat{\boldsymbol{\varepsilon}}_{pj} = \mathbf{y} - \hat{\mathbf{y}} + \hat{f}(\mathbf{x}_j)$$

where the partial residuals for covariate j are the sum of the residuals and the fitted values for covariate j.

An **unbiased estimator of $\boldsymbol{\sigma}^2$** is:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_i (y_i - \hat{y}_i)^2 = \frac{\text{RSS}}{n-p}$$

where $n-p$ is called the *degrees of freedom for error*, df_{err}.

The *degrees of freedom for the model*, df_{mod} is the number of parameters in the model, which in this case is p. Therefore,

$$\text{df}_{\text{err}} = n - \text{df}_{\text{mod}}.$$

QR decomposition

In practice, the formulation above to compute $\hat{\boldsymbol{\beta}}$ is not particularly computationally efficient. Therefore, since,

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2,$$

the value will be unchanged if $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is rotated. This provides the basis for a practical method for finding $\hat{\boldsymbol{\beta}}$, which is often used in practice e.g. for programming in `R` for computational efficiency. Any real matrix $\mathbf{X}$ can always be decomposed,

$$\mathbf{X} = \mathbf{Q}\begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = \mathbf{Q_f R}$$

where $\mathbf{R}$ is a $p \times p$ upper triangular matrix and $\mathbf{Q}$ is an $n \times n$ orthogonal matrix, the first p of which form $\mathbf{Q_f}$. Applying $\mathbf{Q}^\top$ to $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ gives us:

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = ||\mathbf{Q}^\top \mathbf{y} - \mathbf{Q}^\top \mathbf{X}\boldsymbol{\beta}||^2 = \left|\left|\mathbf{Q}^\top \mathbf{y} - \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} \boldsymbol{\beta}\right|\right|^2$$

Writing

$$\mathbf{Q}^\top \mathbf{y} = \begin{bmatrix} \mathbf{f} \\ \mathbf{r} \end{bmatrix},$$

where $\mathbf{f}$ is vector of dimension p and hence $\mathbf{r}$ is a vector of dimension $n - p$ gives us:

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = \left|\left|\begin{bmatrix} \mathbf{f} \\ \mathbf{r} \end{bmatrix} - \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} \boldsymbol{\beta}\right|\right|^2 = ||\mathbf{f} - \mathbf{R}\boldsymbol{\beta}||^2 + ||\mathbf{r}||^2$$

The length of $\mathbf{r}$ does not depend on $\boldsymbol{\beta}$, while $||\mathbf{f} - \mathbf{R}\boldsymbol{\beta}||^2$ can be reduced to zero by choosing $\boldsymbol{\beta}$ so that $\mathbf{R}\boldsymbol{\beta}$ equals $\mathbf{f}$.

Hence,

$$\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{f}$$

is the least squares estimator of $\boldsymbol{\beta}$.

We will not develop this further here, see Wood (2017) for further details.

Maximum likelihood

An alternative method to estimate $\hat{\boldsymbol{\beta}}$ is through using maximum likelihood. Under the assumption of a normal distribution for the errors, the maximum likelihood estimate of the parameters can be found by maximising the following likelihood function.

We have assumed that $\boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathbf{I_n})$, which is equivalent to assuming $\mathbf{y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I_n})$. Therefore,

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2}\sigma^{-n}\exp(-||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2/2\sigma^2).$$

A direct extension of this is the situation where the data do not meet the constant variance assumption, and may not even be independent i.e. the model

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathbf{y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}\sigma^2)$$

where $\mathbf{V}$ is any positive definite matrix. In this case the likelihood for $\boldsymbol{\beta}$ is:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n |\mathbf{V}|}} e^{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/(2\sigma^2)}$$

and if $\mathbf{V}$ is known then the maximum likelihood estimator of $\boldsymbol{\beta}$ can be computed.

The likelihood approach can be taken further. If $\mathbf{V}$ depends on unknown parameters then these too can be estimated by maximum likelihood estimation and this is what is done in linear mixed modelling (see Section 3.3).

In a similar way to the expressions above, the least squares estimates of $\boldsymbol{\beta}$ can be found by minimising:

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

and the weighted least squares formulation is given by:

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

In vector-matrix form we have:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{y}$$

and

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}\mathbf{y}.$$

Please revise assessing model goodness of fit, R^2(adj), model diagnostic checking (e.g. plots of residuals versus fits) and inference for model coefficients (e.g. testing H_0: $\beta_j = 0$, and producing associated confidence intervals).

Model comparison

It is usually a good idea to avoid over-complicated models (i.e. those that are dependent on predictor variables which do not provide additional information), to produce models which are efficient and straight forward to interpret. Several approaches to model selection are based on hypothesis tests about model covariates. For example:

An F-test:

Under the null hypothesis that the simpler model (Model 0) is appropriate:

$$\mathrm{F} = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(\mathrm{df}_0 - \mathrm{df}_1)}{\mathrm{RSS}_1/(\mathrm{df}_1)} \sim \mathrm{F}_{(\mathrm{df}_0 - \mathrm{df}_1, \mathrm{df}_1)}.$$

We reject the null hypothesis for small p-values and conclude that the more complex model provides additional useful information, i.e we reject the simpler model for the more complex one if $\mathrm{F} > \mathrm{F}_{(\mathrm{df}_0 - \mathrm{df}_1, \mathrm{df}_1)}$ at some chosen significance level α.

An alternative method to hypothesis testing is to try to find the model that gets as close as possible to the true model, rather than to find the simplest model that is consistent with the data. Selecting models in order to minimise information criteria (e.g. Akaike's Information Criterion (AIC)) is one way of trying to do this - please revise these ideas.

Exercise: Unbiased estimators
For a linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \quad i = 1, \ldots, n, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}),$$

show that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$ (the vector of parameters) and derive an expression for the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. (Hint: $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, for an unbiased estimator).

Exercise: Degrees of freedom for a model

In the case of a linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_p x_{pi} + \varepsilon_i, \quad i = 1, \ldots, n,$$

show that the trace (the sum of the diagonal entries) of the (so called) hat or projection matrix is $p + 1$ i.e. the trace (tr) of the hat/projection matrix ($\mathbf{H}$) is equal to the number of parameters (degrees of freedom for the model). *It will be helpful to use the fact that tr(*$\mathbf{AB}$*) = tr(*$\mathbf{BA}$*), where* $\mathbf{A}$ *and* $\mathbf{B}$ *are matrices.*

2.2 Bayesian linear model

In this section we will quickly revise the Bayesian linear model, i.e. we assume the linear regression model

$$y_i | \boldsymbol{\beta} \sim \mathsf{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2) \text{ i.i.d.}$$

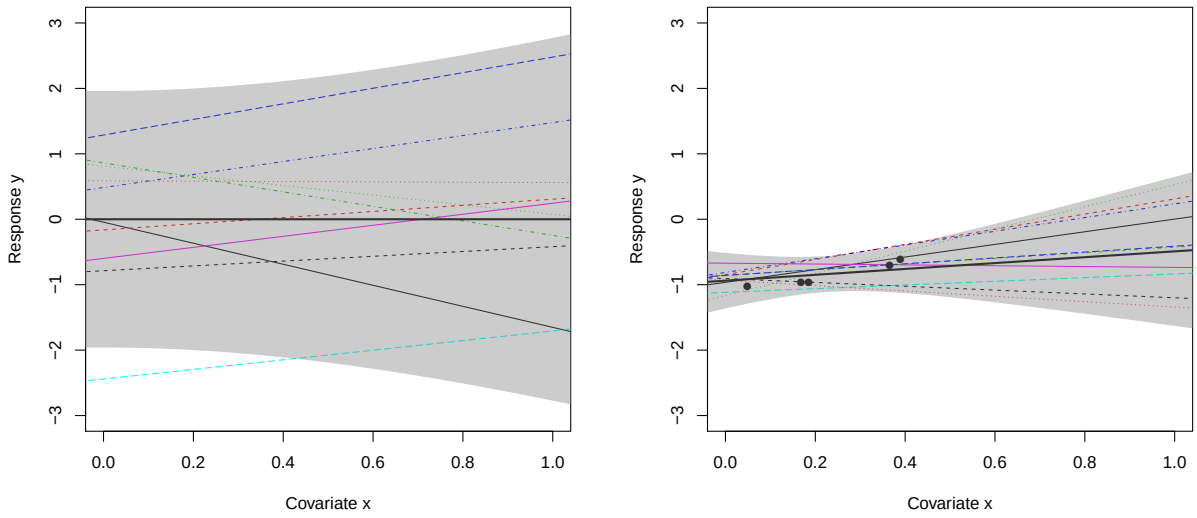
or equivalently,

$$\mathbf{y} | \boldsymbol{\beta} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}).$$

Assume for now that σ^2 is known and just place a Gaussian prior on $\boldsymbol{\beta}$, i.e.

$$\boldsymbol{\beta} \sim \mathsf{N}(\mathbf{0}, \tau^2 \mathbf{I}).$$

We can write the p.d.f. of the posterior distribution of $\boldsymbol{\beta}$ as



(a) Samples from the prior distribution.

(b) Data and samples from the posterior distribution.

Figure 1: Draws from the prior distribution (a) and the posterior distribution (b) of a Bayesian linear model. The bold line corresponds to the mean, the shaded area corresponds to pointwise 95% credible intervals.

$$
\begin{aligned}
f(\boldsymbol{\beta}|y_1, \ldots, y_n) &\propto \underbrace{\left(\prod_{i=1}^{n} f(y_i|\boldsymbol{\beta})\right)}_{\text{Likelihood}} \cdot \underbrace{f(\boldsymbol{\beta})}_{\text{prior}} \\
&= \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right)\right) \cdot \left(\frac{1}{\sqrt{2\pi\tau^2}}\right) \exp\left(-\frac{\sum_{j=1}^{p} \beta_j^2}{2\tau^2}\right).
\end{aligned}
$$

Collecting terms, taking logs and keeping only terms involving $\boldsymbol{\beta}$ yields the log-posterior density

$$
\log f(\boldsymbol{\beta}|y_1, \ldots, y_n) = \text{const} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 - \frac{1}{2\tau^2} \sum_{j=1}^{p} \beta_j^2,
$$

which is, up to a multiplicative constant, the objective function used in ridge regression with $\lambda = \frac{\sigma^2}{\tau^2}$.

One can show (by completing the square) that the posterior distribution of $\boldsymbol{\beta}$ is

$$
\boldsymbol{\beta}|y_1, \ldots, y_n \sim \mathsf{N}\left(\left(\mathbf{X}^\top\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\right)^{-1} \mathbf{X}^\top\mathbf{y}, \left(\mathbf{X}^\top\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\right)^{-1}\right).
$$

Thus the ridge regression estimate $\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$ is the Bayesian maximum-a-posteriori (MAP) estimate of $\boldsymbol{\beta}$.

Figure 1 illustrates this idea of Bayesian inference for a linear model with design matrix $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$. Panel (a) shows ten draws from the prior distribution, whereas panel (b) shows draws from the posterior distribution given the data.

2.3 Generalised linear models (GLMs)

GLMs extend a linear model to the situation where the response variable is a distribution other than the normal, but is a member of the class known as the exponential family of distributions.

A generalised linear model (GLM) can be written in the form:

$$g(\mathbb{E}(Y_i)) = \beta_0 + \beta_1 x_{1i} + \ldots \beta_m x_{mi}$$

or in vector-matrix notation:

$$g(\mathbb{E}(\mathbf{y})) = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}, \qquad g(\mathbb{E}(Y_i)) = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

where, g is a smooth monotonic link function and describes how the mean response is linked to the covariates through the linear predictor:

$$\eta_i = g(\mu_i)$$

$$\mu_i = \mathbb{E}(Y_i), \quad y_i \sim \mathrm{EF}(\mu_i, \phi)$$

and EF is an exponential family distribution with scale parameter ϕ.

In principle, any continuous and differentiable function which is monotonic can be used as the link function but there are some choices which are common and convenient for standard GLMs. See the APTS course - *Statistical Modelling* for revision on exponential families.

In order to estimate the parameters $\boldsymbol{\beta}$ of the GLM we use maximum likelihood estimation.

Maximum likelihood estimation for a set of parameters $\boldsymbol{\beta}$ based on data $\mathbf{y}$:

- $L(\boldsymbol{\beta}; \mathbf{y}) \propto \prod_{i=1}^{n} f(y_i, \boldsymbol{\beta})$
- $\ell(\boldsymbol{\beta}; \mathbf{y}) = \log_{\mathrm{e}}(L(\boldsymbol{\beta}; \mathbf{y}))$
- Usually, $\hat{\boldsymbol{\beta}}$ is obtained by differentiating the log-likelihood wrt each of the parameters and setting equal to zero:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_t} = 0 \quad \text{for } t = 1, \ldots, m.$$

Often the solutions are obtained numerically and likelihood estimates are maxima if the Hessian matrix (matrix of second order derivatives) is negative definite.

Sometimes we can maximise the likelihood analytically and find an exact solution for the MLE $\hat{\boldsymbol{\beta}}$ but the Gaussian GLM is the only common case where this is possible. Typically, we must use numerical optimization. This optimization can be shown to be equivalent to iteratively reweighted least squares (IRWLS).

In order to perform inference for GLMs, the terminology of deviance is required:

$$\text{Deviance} = D = 2\log\lambda = 2(\ell(\hat{\boldsymbol{\beta}}_{\boldsymbol{max}}; \mathbf{y}) - \ell(\hat{\boldsymbol{\beta}}; \mathbf{y})).$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{max}}$ are the estimated parameters for the saturated (full) model.

Models such as the normal distribution and gamma distribution have a scale parameter ϕ and for these it is better to use the scaled deviance. This can be written as $D^* = D/\phi$.

To compare a more complex model 1 to a smaller nested model 0, the difference in the scaled deviances $D_0^* - D_1^*$ is asymptotically χ^2 with degrees of freedom equal to the difference in the number of identifiable parameters in the two models. That is:

$$D_0^* - D_1^* \sim \chi^2(1 - \alpha; p_1 - p_0).$$

The simpler model is rejected in favour of the more complex model if:

$$D_0^* - D_1^* > \chi^2(1 - \alpha; p_1 - p_0)$$

where α is the chosen significance level, and p_1 and p_0 are the number of parameters in Models 1 and 0 respectively.

As with standard linear models, it is important to check the adequacy of the assumptions that we are making. These checks are still based on the residuals, the difference between the observed and the fitted values from the GLM. However, since the variance of the response is not constant for most GLMs they have to be modified to enable them to be used in a similar way to residuals from a Gaussian linear model. Pearson residuals and deviance residuals are reasonable choices.

3 Smoothing

The first model of interest that we introduced in Section 1.1 was to allow flexibility in the mean, and we will do this through using smoothing. In this course we will focus on smoothing within regression models. However, in this preliminary material we will introduce some of the ideas and concepts for smoothing through looking at density estimation - a straight forward example to introduce the ideas (and that can also be useful for fitting mixture models).

3.1 Density estimation

The simple idea of density estimation is to place a kernel function, which in fact is itself a density function, on top of each observation and average these functions.

A probability density function is a key concept through which variability can be expressed precisely. In statistical modelling its role is often to capture variation sufficiently well, within a model where the main interest lies in structural terms such as regression coefficients. However, there are some situations where the shape of the density function itself is the focus of attention. The example below illustrates this.

Example: Aircraft data

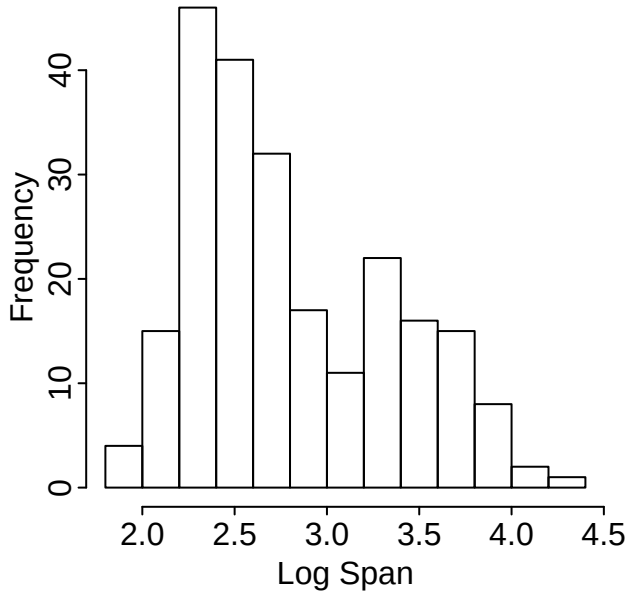
These data record six characteristics of aircraft designs which appeared during the twentieth century. The variables are:

`Yr`	year of first manufacture
`Period`	a code to indicate one of three broad time periods
`Power`	total engine power (kW)
`Span`	wing span (m)
`Length`	length (m)
`Weight`	maximum take-off weight (kg)
`Speed`	maximum speed (km/h)
`Range`	range (km)

The data are available in the `sm` package for `R` through the object `aircraft`:

```
library(sm)
names(aircraft)
```

A brief look at the data suggests that the six measurements on each aircraft should be expressed on the log scale to reduce skewness. Span is displayed on a log scale below, for Period 3 which corresponds to the years after the Second World War. The pattern of variability shown in the histogram exhibits some skewness. There is perhaps even a suggestion of a subsidiary mode at high values of log span, although this is difficult to evaluate.



A simple density estimate

The histogram is a very familiar object. It can be written as

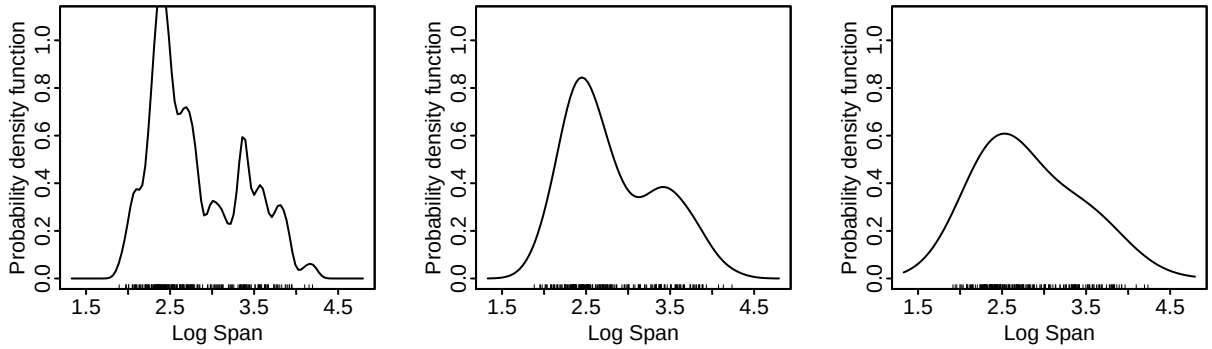
$$\tilde{f}(y) = \sum_{i=1}^{n} I(y - \tilde{y}_i; h),$$

where $y_1, \ldots, y_n$ denote the observed data, $\tilde{y}_i$ denotes the centre of the interval in which y_i falls and $I(z; h)$ is the indicator function of the interval $[-h, h]$. (Notice that further scaling would be required to ensure that $\tilde{f}$ integrates to 1.)

The form of the construction of $\tilde{f}$ highlights some features which are open to criticism if we view the histogram as an estimator of the underlying density function. Firstly the histogram is not smooth, when we expect that the underlying density usually will be. Secondly, some information is lost when we replace each observation y_i by the bin mid-point $\tilde{y}_i$. Both of these issues can be addressed by using a density estimator in the form

$$\hat{f}(y) = \frac{1}{n} \sum_{i=1}^{n} w(y - y_i; h),$$

where w is a probability density, called here a *kernel function*, whose variance is controlled by the *smoothing parameter* h. The middle panel in the plots below shows the effects of doing this with the aircraft data. Large changes in the value of the smoothing parameter have large effects on the smoothness of the resulting estimates, as the left and right hand plots below illustrate.

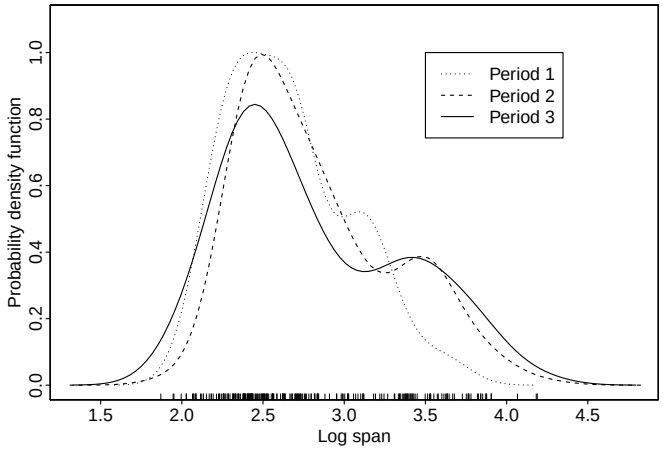


Exercise: The effect of the smoothing parameter

To experiment with density estimates, the code below should launch a new window with interactive controls. Try altering the smoothing parameter through the slider. Does this help you assess whether the subsidiary mode is a genuine feature or an artefact of random variation?

```
library(rpanel)
library(tkrplot)
library(sm)
y <- log(aircraft$Span[aircraft$Period == 3])
sm.density(y, panel = TRUE)
```

One advantage of density estimates is that it is a simple matter to superimpose these to allow different groups to be compared. Here the groups for the three different time periods are compared. It is interesting that the 'shoulder' appears in all three time periods.



Simple properties of density estimates

Without any real restriction, we can assume that the kernel function can be written in the simple form $w(y - y_i; h) = \frac{1}{h} w\left(\frac{y-y_i}{h}\right)$. The mean of a density estimator can then be written as

$$\mathbb{E}\Big\{\hat{f}(y)\Big\} = \int \frac{1}{h} w\left(\frac{y-z}{h}\right) f(z) dz = \int w(u) f(y-hu) du,$$

where the last expression simply involves a change of variable $u = \frac{y-z}{h}$. A Taylor series expansion of the term involving f in the last expression gives

$$f(y-hu) = f(y) - huf'(y) + \frac{1}{2}h^2u^2f''(y) + o(h^2)$$

and, on insertion into the expression for the mean, this produces the approximation

$$\mathbb{E}\Big\{\hat{f}(y)\Big\} \approx f(y) + \frac{h^2}{2}\sigma_w^2 f''(y),$$

where we assume that the kernel function is symmetric so that $\int uw(u)du = 0$, and where σ_w^2 denotes the variance of the kernel, namely $\int u^2 w(u)du$.

The variance of the density estimate can be written as

$$\begin{aligned} \mathrm{Var}\{\hat{f}(y)\} &= \frac{1}{n}\mathrm{Var}\left\{\frac{1}{h}w\left(\frac{y-Y}{h}\right)\right\} \\ &= \frac{1}{n}\left\{\mathbb{E}\left\{\left[\frac{1}{h}w\left(\frac{y-Y}{h}\right)\right]^2\right\} - \mathbb{E}\left\{\frac{1}{h}w\left(\frac{y-Y}{h}\right)\right\}^2\right\}. \end{aligned}$$

A similar change of variable and Taylor series expansion produces the approximation

$$\mathrm{Var}\{\hat{f}(y)\} \approx \frac{1}{nh} f(y)\,\alpha(w),$$

where $\alpha(w) = \int w^2(u)du$.

These expressions capture the essential features of smoothing. In particular, bias is incurred and we can see that this is controlled by f'', which means that where the density has peaks and valleys the density estimate will underestimate and overestimate respectively. This makes intuitive sense.

A useful global measure of performance is the *mean integrated squared error* (MISE) which balances squared bias and variance.

$$\begin{aligned} \mathrm{MISE}(\hat{f}) &= \mathbb{E}\Big\{\int [\hat{f}(y) - f(y)]^2 dy\Big\} \\ &= \int \Big[\mathbb{E}\Big\{\hat{f}(y)\Big\} - f(y)\Big]^2 dy + \int \mathrm{Var}\{\hat{f}(y)\}dy. \end{aligned}$$

3.2 Basis function approaches

When we consider nonparametric regression and (generalised) additive models, kernel weight functions (as introduced in section 3.1), through local regression are one approach that can be used to estimate the smooth fucntion for a covariate relationship with a response (using the data). An alternative approach is to use the idea of basis functions, and so we will introduce/revise the mathematical terminology and notation here.

A basis

Consider the following general polynomial:

$$Y_i = f(x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i.$$

We say that the space of polynomials of order p and below, contains f. This is an example of a *basis*.

A *basis* defines the space of functions of which f (or a close approximation to it) is an element.

Choosing a *basis* amounts to choosing some basis functions which will be treated as completely known. The basis functions for a third order polynomial would be: $1, x, x^2, x^3$ and hence the *basis* is:

$$b_0(x) = 1, b_1(x) = x, b_2(x) = x^2, b_3(x) = x^3.$$

Polynomial regression can be a useful tool if a polynomial of very low order yields a sufficient fit to the data. However, polynomial regression is not very well suited for modelling more complex relationships as each basis function acts "globally" rather than "locally". Fitting data well in one part of the sample space can create artefacts elsewhere.

An alternative approach is to use a set of basis functions which are more local in their effects. Polynomial splines are the most popular such model. Polynomial spline models are based on piecewise polynomial models of low order, and we will explore a truncated power basis, B-splines, fourier basis and p-splines as part of this course.

Basis function approaches are not based on local weights, but based on expanding the design matrix used in linear regression. To fix notation, we quickly state the simple linear regression model

$$\mathbb{E}(Y_i) = f(x_i) = \beta_0 + \beta_1 x_i \qquad \text{for } i = 1, \ldots, n,$$

or equivalently, in matrix-vector notation,

$$\mathbb{E}(\mathbf{y}) = \mathbf{B}\boldsymbol{\beta} \qquad \text{with } \mathbf{y} = (Y_1, \ldots, Y_n)^\top \text{ and } \mathbf{B} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Basis function approaches effectively consist of introducing functions of $\mathbf{x}$ (other than just the identity) into the design matrix $\mathbf{B}$.

One key advantage of basis-expansion methods is that we can then estimate $\boldsymbol{\beta}$ using the same techniques as used in multiple linear regression, i.e. the least-squares estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{y}$$

for a design matrix $\mathbf{B}$ of basis functions rather than $\mathbf{X}$ of covariates.

Exercise: Basis functions
Write down the basis functions for the following models:

- $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i + \beta_{11}(x_i - 0.6)_+ + \varepsilon_i, \quad i = 1, \ldots, n;$
- $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i + \beta_{11}|x_i - \kappa_1| + \varepsilon_i, \quad i = 1, \ldots, n.$

Note: $(x - 0.6)_+$ refers to the positive part of the function and hence this is zero for values of $x < 0.6$, and κ_1 is known.

We will start the course by considering the ideas of nonparametric regression, estimating $f(x)$ in the equation:

$$Y_i = f(x_i) + \varepsilon_i$$

using both kernel and spline based methods.

The two key issues here are:

- how to do the smoothing? (E.g. kernel/spline based methods) and,
- how much to smooth? (I.e. considering the bias/variance trade-off).

3.3 Mixed models

The idea and concepts of random effects can be used to introduce methods of automatically selecting the level of smoothing for smooth functions in regression models.

So let's revise the key ideas for mixed models.

Mixed models contain:

- **Fixed effects**: unknown parameter(s) that we are specifically interested in estimating from the data;
- **Random effects**: random variables where we try to estimate the parameters that describe the distribution of the random effects.

For a mixed effect model with normal errors:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

We assume that the random effects $\boldsymbol{\gamma} \sim \mathsf{N}(0, \mathbf{D})$, $\boldsymbol{\varepsilon} \sim \mathsf{N}(0, \sigma^2 \mathbf{I})$.

This gives us that:

$$\text{Var}(\mathbf{y}) = \text{Var}(\mathbf{Z}\boldsymbol{\gamma}) + \text{Var}(\boldsymbol{\varepsilon}) = \mathbf{Z}\mathbf{D}\mathbf{Z}^\top + \sigma^2\mathbf{I}$$

and hence,

$$\mathbf{y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{D}\mathbf{Z}^\top + \sigma^2\mathbf{I}).$$

Standard maximum likelihood is one method of estimation to find $\boldsymbol{\beta}$, σ^2 and $\mathbf{D}$, with estimation of $\mathbf{V}$ indirectly referred to as variance component estimation. However, maximum likelihood estimation of variance components tends to underestimate them. Therefore, an alternative approach to estimation is Restricted Maximum Likelihood (REML), which involves maximizing the likelihood of linear combinations of the elements of $\mathbf{y}$ that do not depend on the fixed parameters $\boldsymbol{\beta}$ and is a bias reducing alternative.

See the APTS courses *Statistical Modelling* and *Statistial Inference* for more revision here.

An additive model can be represented as a mixed model with a variance component controlling the amount of smoothing for each additive component - we will expand on these ideas within the course to enable automatic selection of smoothing parameters.

4 Quantile regression

The second model of interest in the course is where we allow flexibility in the response being modelled. In this course we will consider the ideas of quantile regression as outlined in Section 1.2. To motivate this, consider the example below:

Example: Effect of age on obesity in the US - part 1

This example explores data on body mass index (BMI) and age from the United States. The example fits a series of models. Firstly,

- a linear regression for the relationship between mean BMI and age;
- a quantile regression to explore the 90% and 98% quantiles of the conditional distribution.
  - To find the 90% and 98% quantile of the conditional distribution of the BMI given Age we fit two quantile regression models using the function `rq` from `quantreg`.

Example `R` code is:

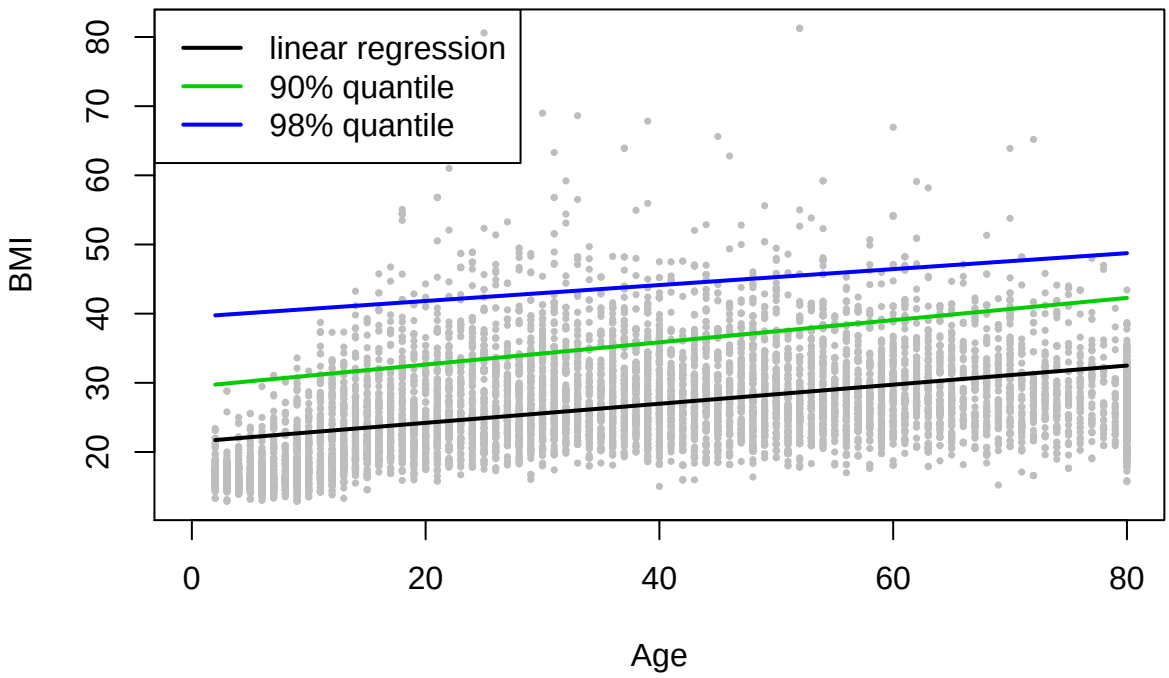
```
library(NHANES)
library(quantreg)

plot(BMI~Age, data=NHANES, col="grey", pch=16, cex=0.5)
newdata <- data.frame(Age=seq(2,80,len=500))

model <- lm(BMI~Age, data=NHANES)
lines(newdata$Age, predict(model, newdata), col=1, lwd=2)

model <- rq(BMI~Age, data=NHANES, tau=0.9)
lines(newdata$Age, predict(model, newdata), col=3, lwd=2)

model <- rq(BMI~Age, data=NHANES, tau=0.98)
lines(newdata$Age, predict(model, newdata), col=4, lwd=2)
legend("topleft",col=c(1,3,4), lwd=2, c("linear regression",
                                        "90% quantile", "98% quantile"))
```



In their vanilla form, quantile regression models are run separately for each quantile. This can however sometimes lead to a problem with estimated quantiles crossing.

This can be avoided by estimating the entire conditional distribution in one go.

Linear programming algorithms are used for estimation here and these concepts will be revised/introduced throughout the course.

Example: Effect of age on obesity in the US - part 2

In this second part, the example explores smooth functions for the relationship between data on body mass index (BMI) and age from the United States. Specifically,

- a nonparametric regression using B-splines to estimate the smooth function;
- a nonparametric quantile regression using B-splines to explore the 90% and 98% quantiles of the conditional distribution

Example R code is:

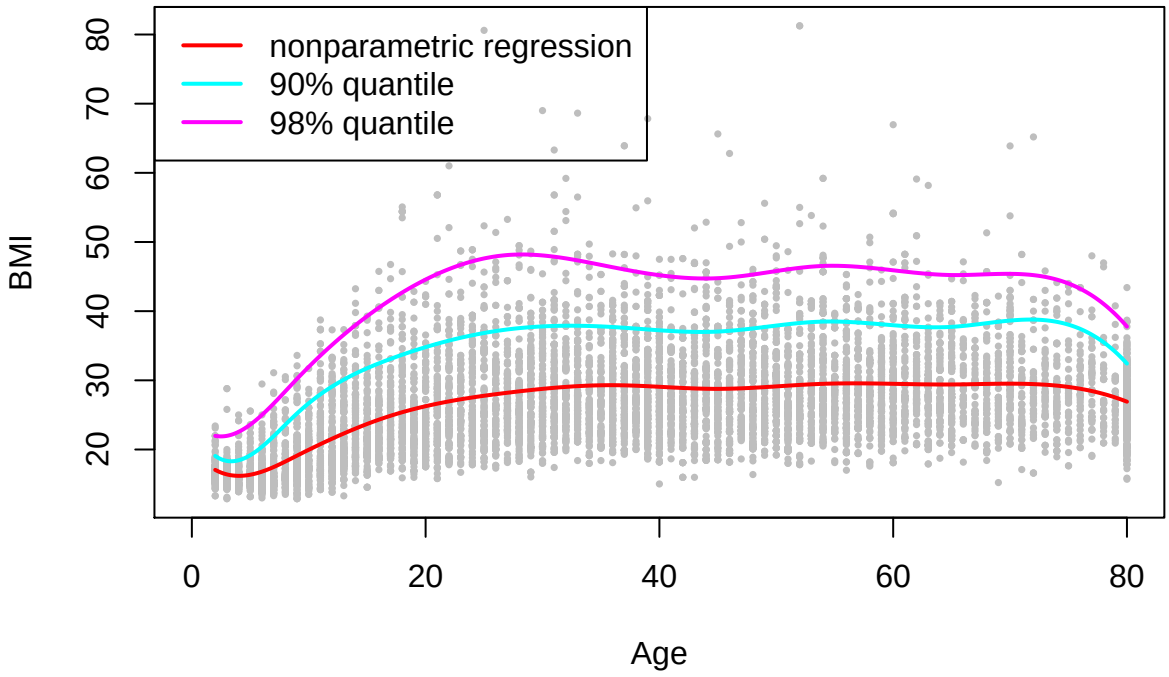
```
library(NHANES)
library(splines)
library(quantreg)

plot(BMI~Age, data=NHANES, col="grey", pch=16, cex=0.5)
newdata <- data.frame(Age=seq(2,80,len=500))
```

```
model <- lm(BMI~bs(Age, df=10), data=NHANES)
lines(newdata$Age, predict(model, newdata), col=2, lwd=2)

model <- rq(BMI~bs(Age, df=10), data=NHANES, tau=0.9)
lines(newdata$Age, predict(model, newdata), col=5, lwd=2)

model <- rq(BMI~bs(Age, df=10), data=NHANES, tau=0.98)
lines(newdata$Age, predict(model, newdata), col=6, lwd=2)
legend("topleft",col=c(2,5,6), lwd=2, c("nonparametric regression",
                                        "90% quantile", "98% quantile"))
```



Which of these models do you think is a useful representation for the pattern/relationship?

This course will introduce and develop the theoretical details and concepts for the models fitted in the example above (including B-splines and quantile regression), and extend these ideas for the situation of multiple covariates.

5 Broad concepts

The simple case of density estimation and the introduction to the area of quantile regression here highlight features and issues which are common to a wide range of problems involving the estimation of functions, relationships or patterns which are *nonparametric* but *smooth*, and evident over different quantiles of the response distribution.

The term nonparametric is used in this context to mean that the relationships or patterns of interest cannot be expressed in specific formulae which involve a fixed number of unknown parameters. This means that the parameter space is the space of functions, whose dimensionality is infinite. This takes us outside of the standard framework for parametric models and the main theme of the course will be to discuss how we can do this while producing tools which are highly effective for modelling and analysing data from a wide variety of contexts and exhibiting a wide variety of structures.

On a side note, the term nonparametric is sometimes used in the narrower setting of simple statistical methods based on the ranks of the data, rather than the original measurements. This is not the sense in which it will be used here.

6 R packages

The following R packages will be required in the labs if you intend to use your own laptop.

- gamlss
- ggplot2
- mgcv
- quantreg
- rpanel
- splines

7 Further reading

A variety of texts on flexible regression:

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall: London.

Bowman, A.W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. OUP: Oxford.

Koenker, R. (2005). *Quantile Regression* (Econometric Society Monographs). Cambridge: Cambridge University Press.

Koenker, R., Chernozhukov, V., He, X. & Peng, L. (2017). *Handbook of Quantile Regression*. Chapman and Hall/CRC.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric Regression*. CUP: Cambridge.

Wood, S. (2017). *Generalized additive models: an introduction with* `R` *SECOND EDITION*. CRC Press, Taylor & Francis Group, Boca Raton, FL.

8 Solutions

Exercise: Unbiased estimators

For $\hat{\boldsymbol{\beta}}$ to be an unbiased estimator of $\boldsymbol{\beta}$ we require $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}) \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{E}(\mathbf{y}) \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathrm{Cov}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}) \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{Cov}(\mathbf{y})((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^\top \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2 \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2
\end{aligned}$$

Exercise 2: Degrees of freedom for a model

$$\begin{aligned}
\mathrm{tr}(\mathbf{H}) = \mathrm{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) &= \mathrm{tr}(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}) \\
&= \mathrm{tr}(\mathbf{I}_{p+1}) = p+1
\end{aligned}$$

Exercise 3: Basis functions

- $1, x, (x-0.6)_+$
- $1, x, |x-\kappa_1|$