CHAPTER 11

Extensions of Generalized Linear Models

This final chapter introduces alternatives to maximum likelihood (ML) and Bayes for fitting linear and generalized linear models (GLMs). We also present an extension of the GLM that permits an additive predictor in place of the linear predictor. A complete exposition of these topics is beyond the scope of this book. We aim here merely to present a brief overview and give you references for further study.

Section 11.1 presents alternative ways to estimate model parameters. For the linear model, *M-estimation* methods minimize a function of the residuals, the sum of squared residuals being one special case. Some such estimates are more robust than least squares, because they are less affected by severe outliers or by contamination of the data. *Regularization methods* modify ML to give sensible answers in situations that are unstable because of causes such as collinearity. For the GLM, the *penalized likelihood* regularization method modifies the log-likelihood function by adding a penalty term, resulting in estimates that tend to have smaller variance than ML estimators.

Regularization methods are especially useful when the number p of model parameters is very large. Such datasets are common in genomics, biomedical imaging, functional magnetic resonance imaging, tomography, signal processing, image analysis, market basket data, and portfolio allocation in finance. Sometimes p is even larger than n. Section 11.2 discusses the fitting of GLMs with high-dimensional data, focusing on identifying the usually small subset of the explanatory variables that are truly relevant for modeling $E(y)$.

Another extension of the ordinary GLM replaces the linear predictor by smooth functions of the explanatory variables. Section 11.3 introduces generalizations of the GLM that do this, such as the *generalized additive model*, or that have structure other than modeling the mean response with a linear predictor, such as *quantile regression*

Foundations of Linear and Generalized Linear Models, First Edition. Alan Agresti.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

for modeling quantiles of the response distribution and *nonlinear regression* when the response mean is a nonlinear function of parameters.

11.1 ROBUST REGRESSION AND REGULARIZATION METHODS FOR FITTING MODELS

For an ordinary linear model with residuals $\{e_i = y_i - \hat{\mu}_i\}$, the least squares method minimizes $\sum_i e_i^2$. The model fit can be severely affected by observations that have both large leverage and a large residual (recall Section 2.5.5). So that such observations have less influence, we could instead minimize a function that gives less weight to large residuals.

11.1.1 M-Estimation for Robust Regression

An alternative function to minimize is $\sum_i \rho(e_i)$ for an objective function $\rho(e_i)$ that is symmetric with a minimum at 0 but with possibly less than a quadratic increase. This approach is called *M-estimation*. Like least squares, it does not require assuming a distribution for y.

In M-estimation, the estimates $\hat{\boldsymbol{\beta}}$ of the parameters $\boldsymbol{\beta}$ in the linear predictor are the solutions to the equations

$$\frac{\partial}{\partial \hat{\beta}_j}\left[\sum_{i=1}^{n} \rho(e_i)\right] = \sum_{i=1}^{n} \frac{\partial \rho(e_i)}{\partial e_i}\frac{\partial e_i}{\partial \hat{\beta}_j} = 0, \quad j = 1, \ldots, p.$$

For the linear model, $\partial e_i / \partial \hat{\beta}_j = -x_{ij}$. The function $\psi(e) = \partial \rho(e)/\partial e$ is called the *influence function*, because it describes the influence of an observation's residual on $\hat{\boldsymbol{\beta}}$. For least squares, the influence increases linearly with the size of the residual. A more robust solution chooses $\rho(e)$ so that $\psi(e)$ is a bounded function.

Let $\boldsymbol{\rho}(\boldsymbol{\beta})$ represent ρ expressed in terms of the population residuals $\{\epsilon_i = y_i - \boldsymbol{x}_i\boldsymbol{\beta}\}$ as $[\rho(y_1 - \boldsymbol{x}_1\boldsymbol{\beta}), \ldots, \rho(y_n - \boldsymbol{x}_n\boldsymbol{\beta})]^{\mathrm{T}}$ and satisfying $E[\partial \boldsymbol{\rho}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}] = \mathbf{0}$. Choosing $\rho(\cdot)$ to be strictly convex ensures that a unique estimate exists. A natural choice is the absolute value metric, $\rho(e_i) = |e_i|$. For the null model, this produces the sample median as the estimate of location. But $\rho(e_i)$ is not then strictly convex, the solution may be indeterminate, and the estimator loses considerable efficiency relative to least squares when the normal linear model is adequate. An alternative is $\rho(e) = |e|^p$ for some $1 < p < 2$, although then the influence function is not bounded.

A compromise approach, suggested by Peter Huber in the early literature on M-estimation, takes $\rho(e) = |e|^2$ for small $|e|$ and takes $\rho(e)$ to be a linear function of $|e|$ beyond that point. A popular implementation takes $\rho(e)$ quadratic for $|e| \leq k\hat{\sigma}$, where $k \approx 1.5$ (proposed by Huber) and $\hat{\sigma}$ is a robust estimate of $\sqrt{\text{var}(\epsilon)}$, such as the median absolute residual for the least squares fit divided by 0.67. Smaller values for k protect against a higher proportion of outlying observations, but at a greater loss of efficiency when the normal linear model truly holds. The value $k = 1.345$ provides

95% efficiency under the normal linear model. Other proposals for robust fitting include one by John Tukey for which the influence function is 0 at large absolute values. This completely removes the influence of large outliers.

For a weight function defined by $w(e) = \psi(e)/e$, the estimating equations for the M-estimates $\hat{\boldsymbol{\beta}}$ are

$$\sum_{i=1}^{n} w(e_i)e_i x_{ij} = 0, \quad j = 1, \ldots, p.$$

Finding the solution requires iterative methods, with initial values such as the least squares estimates. At stage t of the iterative process, the estimating equations correspond to those for the iteratively reweighted least squares solution for minimizing $\sum_i w(e_i^{(t)})e_i^2$. That is, for a model matrix $\boldsymbol{X}$ and with $\boldsymbol{W}^{(t)}$ a diagonal matrix having elements $\{w(e_i^{(t)})\}$, their solution is

$$\hat{\boldsymbol{\beta}}^{(t)} = \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{y}.$$

The asymptotic covariance matrix of the limit $\hat{\boldsymbol{\beta}}$ of this iterative process is

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\frac{E[\psi(\epsilon)]^2}{\{E[\psi'(\epsilon)]\}^2}.$$

Substituting the sample analogs $\{\sum_i[\psi(e_i)]^2\}/n$ for $E[\psi(\epsilon)]^2$ and $[\sum_i \psi'(e_i)]/n$ for $E[\psi'(\epsilon)]$ yields an estimated covariance matrix. Fitting is available in software[1].

11.1.2 Penalized-Likelihood Methods

In fitting GLMs, *regularization methods* modify ML to give sensible answers in unstable situations. A popular way to do this adds a term to the log-likelihood function such that the solution of the modified likelihood equations smooths the ordinary estimates. For a model with log-likelihood function $L(\boldsymbol{\beta})$, we maximize

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - s(\boldsymbol{\beta}),$$

where $s(\cdot)$ is a function such that $s(\boldsymbol{\beta})$ decreases as elements of $\boldsymbol{\beta}$ are smoother in some sense, such as uniformly closer to 0. This smoothing method, referred to as *penalized likelihood*, shrinks the ML estimate toward $\mathbf{0}$. Among its positive features are a reduction in prediction error and existence when the ML estimate is infinite or badly affected by collinearity.

[1]For example, the `rlm` (robust linear modeling) function in the R `MASS` package

A variety of penalized-likelihood methods use the L_q-*norm* smoothing function

$$s(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|^q$$

for some $q \geq 0$ and $\lambda \geq 0$. The explanatory variables should be standardized, as they are treated the same way in the smoothing function, and the degree of smoothing should not depend on the choice of scaling. The response variable should also be standardized, or the intercept should be removed from the smoothing term, because there is no reason to shrink a parameter whose estimate (for the ordinary linear model) is merely the overall sample mean response. The constant λ is called a *smoothing parameter*, because the degree of smoothing depends on it. The choice of λ reflects the bias–variance tradeoff discussed in Section 4.6.2. Increasing λ results in greater shrinkage toward 0 in $\{\hat{\beta}_j\}$ and smaller variance but greater bias.

How well a smoothing method works depends on λ. This is usually chosen by cross-validation. For each λ value in a chosen grid, we fit the model to part of the data and then check the goodness of the predictions for y in the remaining data. With k-fold cross-validation, we do this k times (for k typically about 10), each time leaving out the fraction $1/k$ of the data and predicting those y values using the model fit from the rest of the data. The selected value of λ is the one having the lowest sample mean prediction error for the k runs, for a measure of prediction error such as squared difference between observed and predicted y. We then apply that value with the penalized-likelihood method for all the data.

At each λ, the sample mean prediction error is a random variable. An alternative choice for λ uses a *one standard error rule*, in which the chosen λ has mean prediction error that is one standard error above the minimum, in the direction of greater regularization. Such a choice may be less likely to overfit the model.

Penalized-likelihood estimators have Bayesian connections. With prior pdf proportional to $\exp[-s(\boldsymbol{\beta})]$, the Bayesian posterior pdf is proportional to the penalized-likelihood function. The mode of the posterior distribution then equals the penalized-likelihood estimate.

11.1.3 L_2-Norm Penalty: Ridge Regression

Regularization methods that penalize by a quadratic term, such as $s(\boldsymbol{\beta}) = \lambda \sum_j \beta_j^2$, are called L_2-*norm* penalty methods. For normal linear models, the best known such method is *ridge regression*, which finds the value of $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

Equivalently, this solution minimizes $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ subject to $\sum_j \beta_j^2 \leq \lambda^*$, where a 1–1 inverse correspondence holds between λ and λ^*. The solution for the ridge regression estimate is

$$\tilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y},$$

which adds a "ridge" to the main diagonal of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ before inverting it. This modification is helpful when the model matrix is ill-conditioned, such as under collinearity. Adding the ridge makes the matrix invertible, even if $\boldsymbol{X}$ does not have full rank. Since $\tilde{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{y}$, for the ordinary linear model

$$\text{var}(\tilde{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1}.$$

The least squares estimate is the limit of $\tilde{\boldsymbol{\beta}}$ as $\lambda \to 0$. As λ increases, the effect is to shrink the least squares estimate toward $\mathbf{0}$. For example, when explanatory variables are linearly transformed so that $\boldsymbol{X}$ is orthonormal (i.e., $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{I}$), we see that $\tilde{\boldsymbol{\beta}}$ relates to the least squares estimate $\hat{\boldsymbol{\beta}}$ by $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}/(1+\lambda)$. The ridge regression estimate $\tilde{\boldsymbol{\beta}}$ has the form of the Bayesian posterior mean for the normal linear model presented in Section 10.2.2, when the prior mean for $\boldsymbol{\beta}$ is $\mathbf{0}$ and λ here is identified with σ^2/τ^2 in that Bayesian formulation.

11.1.4 L_1-Norm Penalty: The Lasso

Fitting using the L_1-*norm* penalty, for which $s(\boldsymbol{\beta}) = \lambda \sum_j |\beta_j|$, is referred to as the *lasso* ("least absolute shrinkage and selection operator") method (Tibshirani 1996). Equivalently, it maximizes the likelihood function subject to the constraint that $\sum_j |\beta_j| \leq \lambda^*$ for a constant λ^* inversely related to λ. The larger the value of λ, the greater the shrinkage of estimates toward 0. The shrinkage is by a fixed amount, rather than by a fixed proportion as in ridge regression. For λ sufficiently large, this method shrinks some $\hat{\beta}_j$ completely to zero. In constraining $\sum_j |\beta_j| \leq \lambda^*$, the region of acceptable $\{\beta_j\}$ is a region around the origin that is square when $p = 2$. It intersects the contours of the log likelihood, which are elliptical for normal linear models and approximately so for large n with other GLMs, at axes rather than at the interior in which all $\beta_j \neq 0$. Figure 11.1 illustrates. It is informative to plot the penalized estimates as a function

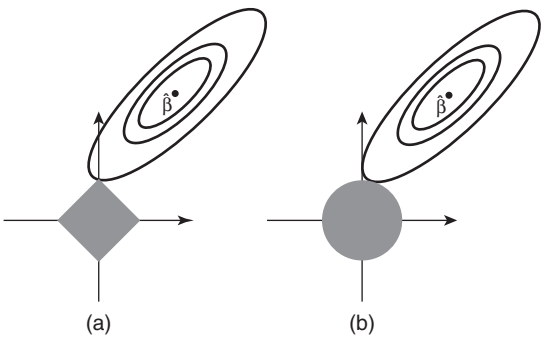

Figure 11.1 Elliptical (or near-elliptical) contours of a GLM log-likelihood function and square contour of the constraint function for the lasso and circular contour of the constraint function for ridge regression. The lasso estimates occur where an ellipse touches the square constraint, often resulting in some $\hat{\beta}_j = 0$. Source: Hastie et al. (2009, p. 71, Figure 3.11), with kind permission from Springer Science+Business Media B.V.

of the permitted value λ^* for $\sum_j |\beta_j|$, to summarize how explanatory variables drop out as λ^* decreases.

Why can shrinking $\{\hat{\beta}_j\}$ toward 0 be effective? In many settings having a large number of explanatory variables, most of them have no effects or very minor effects. An example is genetic association studies, which simultaneously consider each of possibly thousands of genes for the association between the genetic expression levels and the response of interest, such as whether a person has a particular disease. Unless n is extremely large, because of sampling variability the ordinary ML estimates $\{\hat{\beta}_j\}$ tend to be much larger in absolute value than the true values $\{\beta_j\}$. This tendency is exacerbated when we keep only statistically significant variables in a model. Shrinkage toward 0 tends to move $\{\hat{\beta}_j\}$ closer to $\{\beta_j\}$. This is yet another example of the bias–variance tradeoff. Introducing a penalty function results in biased estimates but benefits from reducing the variance.

Penalizing by absolute-value terms makes model fitting more difficult than ridge regression. The estimate of $\boldsymbol{\beta}$ is not linear in $\boldsymbol{y}$, and we need an optimization method to find it. One approach uses the LARS method, to be introduced in Section 11.2.1. A faster method uses coordinate descent, optimizing each parameter separately while holding all the others fixed, and cycling until the estimates stabilize. In a particular cycle t, for explanatory variable x_j, one regresses the residuals $\{y_i - \sum_{k \neq j} \hat{\beta}_k^{(t)} x_{ik}\}$ on $(x_{1j}, \ldots, x_{nj})$ to obtain a value which, when reduced in absolute value by an amount dependent on λ, yields the next approximation for $\hat{\beta}_j$. In practice, this is done for a grid of λ values.

Likewise, finding an estimated covariance matrix for lasso estimators is challenging, especially for the parameters having lasso estimates of 0. Tibshirani (1996) noted that the lasso estimate corresponds to a Bayesian posterior mode for the normal linear model when the independent prior distribution for each β_j is a double-exponential (Laplace) distribution, which has pdf

$$g(\boldsymbol{\beta} \mid \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}}.$$

Each component of this prior distribution has a sharp peak at $\beta_j = 0$. Park and Casella (2007) and Hans (2009) used this result as a mechanism for point and interval estimation of $\{\beta_j\}$.

11.1.5 Comparing Penalized Methods, and Generalizations

Ridge regression, the lasso, and other regularization methods are available in software[2]. A disadvantage of ridge regression is that it requires a separate strategy for finding a parsimonious model, because all explanatory variables remain in the model. By contrast, with the lasso, when λ is large, some $\hat{\beta}_j$ shrink to zero, which can help

[2]In R, the `glmnet` and `ridge` packages and the `lm.ridge` function in the `MASS` package provide ridge regression, and the `glmnet` and `lars` packages provide lasso fits.

with model selection. For a factor predictor, the ordinary lasso solution may select individual indicators rather than entire factors, and the solution may depend on the coding scheme, so an alternative *grouped lasso* should be used. A disadvantage of the lasso is that $\{\hat{\beta}_j\}$ are not asymptotically normal and can be highly biased, making inference difficult. Another disadvantage is that the lasso may overly penalize β_j that are truly large. Which of ridge regression and the lasso performs better in terms of bias and variance for estimating the true $\{\beta_j\}$ depends on their values. When p is large but only a few $\{\beta_j\}$ are practically different from 0, the lasso tends to perform better, because many $\{\hat{\beta}_j\}$ may equal 0. When $\{\beta_j\}$ do not vary dramatically in substantive size, ridge regression tends to perform better.

L_0-*norm* penalty regularization takes $s(\boldsymbol{\beta})$ to be proportional to the number of nonzero β_j. This approach has the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) as special cases. This sounds ideal, but optimization for this criterion is impractical with large numbers of variables; for example, the function minimized may not be convex. A compromise method, *SCAD* (smoothly clipped absolute deviation), starts at the origin $\boldsymbol{\beta} = \mathbf{0}$ like an L_1 penalty and then gradually levels off (Fan and Lv 2010). An alternative *elastic net* uses a penalty function that has both L_1 and L_2 terms (Zou and Hastie 2005). It has both ridge regression and the lasso as special cases. Zou (2006) proposed an *adaptive lasso* that can be better for satisfying an *oracle* property, by which asymptotically the method recovers the correct model and has estimators converging to the parameter values at the optimal rate. It uses an adaptive weighted penalty $\sum_j w_j|\beta_j|$ where $w_j = 1/|\hat{\beta}_j|^\gamma$ for a consistent estimator $\hat{\beta}_j$ such as from least squares, and $\gamma > 0$. This has the effect of reducing the penalty when an effect seems to be large.

11.1.6 Example: House Selling Prices Revisited

In Section 4.7.1 we modeled y = the selling price of a house (in thousands of dollars), using as explanatory variables the size of the house, the property tax bill, whether the home is new, the number of bedrooms, and the number of bathrooms. We now illustrate methods of this section by comparing the least squares fit with other methods, for the simple linear model having all the main effects but no interactions. Adjusted for the other variables, the least squares fit shows strong evidence that the mean selling price increases as the house size increases, as the tax bill increases, and for new houses.

```
------------------------------------------------------------------------------
> attach(Houses) # File Houses.dat at www.stat.ufl.edu/~aa/glm/data
> summary(lm(price ~ size + taxes + new + beds + baths))
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    4.5258     24.4741    0.185    0.8537
size           0.0683      0.0139    4.904  3.92e-06
taxes          0.0381      0.0068    5.596  2.16e-07
new           41.7114     16.8872    2.470    0.0153
beds         -11.2591      9.1150   -1.235    0.2198
baths         -2.1144     11.4651   -0.184    0.8541
------------------------------------------------------------------------------
```

For a robust M-estimation fit, we use the Huber influence function mentioned in Section 11.1.1 with $k = 1.345$ and a robust standard deviation estimate. Summaries of effects are similar to least squares, a notable exception being the effect of *new*. The least squares estimated difference of \$41,711 between the mean selling prices of new and older homes, adjusting for the other variables, decreases to \$27,861.

```
------------------------------------------------------------------------
> library(MASS)
> summary(rlm(price ~ size + taxes + new + beds + baths, psi=psi.huber))
Coefficients: # robust (Huber) fit of linear model
              Value  Std. Error  t value
(Intercept)  11.6233     19.1847   0.6059
size          0.0705      0.0109   6.4533
taxes         0.0341      0.0053   6.3838
new          27.8610     13.2375   2.1047
beds        -16.4034      7.1451  -2.2958
baths         3.9534      8.9873   0.4399
------------------------------------------------------------------------
```

An alternative parametric check fits the model assuming a gamma distribution for y, which naturally accounts for larger variability in selling prices when the mean is larger. The estimated effect of a new home is also then considerably weaker. The change in the *new* effect for these two fits, relative to least squares, is mainly caused by observation 64 in the data file. This observation, which had a relatively low selling price for a very large house that was not new, was an outlier and influential for least squares but not unusual for the gamma model.

```
------------------------------------------------------------------------
> summary(glm(price ~ size + taxes + new + beds + baths,
+             family = Gamma(link=identity)))
Coefficients: # fit of gamma GLM with identity link
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   19.1859     13.7759    1.393    0.1670
size           0.0617      0.0125    4.929   3.5e-06
taxes          0.0378      0.0051    7.475   4.0e-11
new           22.6704     19.3552    1.171    0.2444
beds         -19.2618      6.3273   -3.044    0.0030
baths          9.5825      6.4775    1.479    0.1424
------------------------------------------------------------------------
```

When we implemented the lasso in R with `glmnet`, which operates on the standardized variables, the smoothing parameter value $\lambda = 8.3$ gave the minimum value of cross-validated mean squared error. This fit is not much different from the robust fit but removes *beds* and *baths*, the two predictors that were not significant in the least squares fit. For contrast, we show the coefficients obtained with the much larger value of $\lambda = 23.1$ suggested by the one standard error rule. That fit also removes *new* and has diminished effects of *size* and *taxes*. The first panel of Figure 11.2 shows how the lasso estimates change as λ increases (on the log scale). The *new* estimate decreases from the least squares value of 41.7, becoming 0 when $\log(\lambda) \geq \log(21.0) = 3.0$. For the scaling used, the *size* and *taxes* estimates (which are nonzero for much larger λ values) are too small to appear in the figure. To show this more clearly, the second panel of Figure 11.2 shows the estimates for the standardized variables, for which

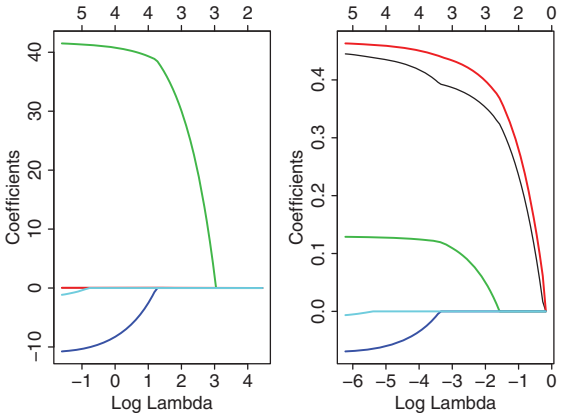

Figure 11.2 Plot of lasso estimates for house selling price data, as function of smoothing parameter $\log(\lambda)$. In the first panel, the least squares estimate of 41.7 for *new* decreases to 0 at $\log(\lambda) = 3.0$. The second panel shows estimates for the standardized variables, from which it is clearer that *size* and *taxes* remain in the model the longest as λ increases. In that panel, the least squares estimates for (size, taxes, new, beds, baths) are (0.45, 0.47, 0.13, −0.07, −0.01), the values of the five curves at the left axis, where λ is essentially 0.

the least-squares estimated effects for (size, taxes, new, beds, baths) are (0.45, 0.47, 0.13, −0.07, −0.01).

```
-------------------------------------------------------------------------
> library(glmnet)
> x <- cbind(size, taxes, new, beds, baths)
> set.seed(1010) # random seed for cross-validation
> cv.glmnet(x,price,alpha=1) # alpha=1 specifies lasso for cross-valid.
$lambda.min # best lambda by 10-fold cross-validation
[1] 8.2883
$lambda.1se # lambda suggested by one standard error rule
[1] 23.06271
> coef(glmnet(x,price, alpha=1, lambda=8.2883))
(Intercept) -5.9947
size         0.0568
taxes        0.0344
new         28.0744
beds         .
baths        .
>  coef(glmnet(x,price, alpha=1, lambda=23.0627))
(Intercept) 22.1646
size         0.0475
taxes        0.0293
new          .
beds         .
baths        .
> fit.lasso <- glmnet(x, price, alpha=1)
> plot(fit.lasso, "lambda")
-------------------------------------------------------------------------
```

For ridge regression, cross-validation suggested using $\lambda = 17.9$. With it, results are not much different from least squares. The fit slightly shrinks the least squares estimates, except for the *new* effect. For $\lambda = 95.3$ from the one standard error rule, the effects of *beds* and *baths* change sign from their least squares values. Keep in mind that for ridge regression and the lasso, results depend greatly on the chosen smoothing parameter λ, and the value chosen for λ in cross-validation will vary considerably according to the seed. One could also report the estimates and standard errors for the standardized variables.

```
------------------------------------------------------------------------
> cv.glmnet(x,price,alpha=0) # alpha=0 specifies ridge regression
$lambda.min     $lambda.1se
[1] 17.85662    [1] 95.2954
> coef(glmnet(x, price, alpha=0, lambda=95.2954))
(Intercept) -4.4871
size         0.0377
taxes        0.0216
new         41.6077
beds         6.4325
baths       16.9838
------------------------------------------------------------------------
```

11.1.7 Penalized Likelihood for Logistic Regression

Penalizing a log-likelihood function need not necessarily result in increased bias. One version actually reduces bias of ML estimators. For most models, the ML estimator $\hat{\boldsymbol{\beta}}$ has bias on the order of $1/n$. Firth (1993) penalized the log likelihood in a way that introduces a small bias into the score function but reduces the bias of $\hat{\boldsymbol{\beta}}$ to order $1/n^2$. For the canonical parameter of an exponential family model, Firth's penalized log-likelihood function uses the determinant of the information matrix $\boldsymbol{\mathcal{J}}$,

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \frac{1}{2}\log|\boldsymbol{\mathcal{J}}|.$$

The penalized likelihood is proportional to the Bayesian posterior distribution resulting from using the Jeffreys prior distribution. Thus, this penalized ML estimator equals the mode of the posterior distribution induced by the Jeffreys prior.

For logistic regression, Firth noted that the ML estimator is biased away from 0, and the bias correction shrinks the estimator toward 0. When the model matrix is of full rank, $\log|\boldsymbol{\mathcal{J}}|$ is strictly concave. Maximizing the penalized log likelihood yields a maximum penalized-likelihood estimate that always exists and is unique. For the null logistic model and a proportion y of successes in n independent Bernoulli trials, it yields as estimate the *empirical logit*, $\log[(ny + \frac{1}{2})/(n - ny + \frac{1}{2})]$. This corresponds to adding $\frac{1}{2}$ to the success and failure counts. Firth's method is especially appealing for the analysis of data that exhibit complete or quasi-complete separation, because then at least one ordinary ML estimate is infinite or does not exist (Section 5.4.2).

11.1.8 Example: Risk Factors for Endometrial Cancer Revisited

Sections 5.7.1 and 10.3.2 described a study about endometrial cancer that analyzed y = histology of 79 cases (0 = low grade, 1 = high grade), with the explanatory variables x_1 = neovasculation (1 = present, 0 = absent), x_2 = pulsatility index of arteria uterina, and x_3 = endometrium height. For the main-effects model

$$\text{logit}[P(y_i = 1)] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

all 13 patients having $x_{i1} = 1$ had $y_i = 1$. So quasi-complete separation occurs, and the ML estimate $\hat{\beta}_1 = \infty$.

```
------------------------------------------------------------------------------
> Endometrial # File Endometrial.dat at www.stat.ufl.edu/~aa/glm/data
   NV PI   EH HG
1   0 13 1.64  0
2   0 16 2.26  0
...
79  0 33 0.85  1
> attach(Endometrial)
> PI2 <- (PI-mean(PI))/sd(PI); EH2 <- (EH-mean(EH))/sd(EH); NV2 <- NV-0.5
> fit.ML <- glm(HG ~ NV2 + PI2 + EH2, family=binomial)
> summary(fit.ML) # ML estimate of NV effect is actually infinite
Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)    7.8411    857.8755    0.009   0.9927
NV            18.1856   1715.7509    0.011   0.9915
PI2           -0.4217      0.4432   -0.952   0.3413
EH2           -1.9219      0.5599   -3.433   0.0006
------------------------------------------------------------------------------
```

Table 10.1 showed Bayes estimates, which with standardized x_2 and x_3 shrink $\hat{\beta}_1$ to 9.1 for quite diffuse normal priors ($\sigma = 10$) and to 1.65 for very informative priors ($\sigma = 1$). The maximum penalized-likelihood estimate for β_1 of 2.93 and the 95% profile penalized-likelihood confidence interval of (0.61, 7.85) shrink the ML estimate $\hat{\beta}_1$ and the ordinary profile likelihood interval of (1.28, ∞) considerably toward 0. Results for the other estimates do not change as much.

```
------------------------------------------------------------------------------
> library(logistf)
> fit.penalized <- logistf(HG ~ NV2 + PI2 + EH2, family=binomial)
> summary(fit.penalized)
Confidence intervals and p-values by Profile Likelihood
               coef se(coef) lower 0.95 upper 0.95   Chisq          p
(Intercept)  0.3080   0.8006    -0.9755     2.7888   0.169  6.810e-01
NV2          2.9293   1.5508     0.6097     7.8546   6.798  9.124e-03
PI2         -0.3474   0.3957    -1.2443     0.4045   0.747  3.875e-01
EH2         -1.7243   0.5138    -2.8903    -0.8162  17.759  2.507e-05
------------------------------------------------------------------------------
```

11.2 MODELING WITH LARGE *P*

High-dimensional data are not well handled by the traditional model-fitting methods presented in this book. In genomics, such applications include classifying tumors by using microarray gene expression or proteomics data or associating protein concentrations with expression of genes or predicting a clinical prognosis by using gene expression data. Generalized linear modeling by ML can be overwhelmed when it needs to detect effects for such applications as differential expression (change between two or more conditions) in many thousands of genes or brain activity in many thousands of locations. We now discuss issues in fitting linear models and GLMs to high-dimensional data in which p is very large, sometimes even with $p > n$. Certain issues are vital yet difficult, such as how to select explanatory variables from an enormous set when nearly all of them are expected to have no effect or a very small effect.

11.2.1 Issues in Variable Selection and Dimension Reduction

In modeling with a very large number of explanatory variables, removing variables that have little if any relevance can ease interpretability and decrease prediction errors. For example, in disease classification, very few of a large number of genes may be associated with the disease. This is reflected by histograms of P-values for testing those effects, which often have appearance similar to the uniform density function that theoretically occurs when the null hypothesis is true. With large p and huge n, ordinary ML fitting may not even be possible and alternative methods may be needed (Toulis and Airoldi 2014). For a binary response, complete or quasi-complete separation often occurs when the number of predictors exceeds a particular point, resulting in some infinite estimates. Even when finite estimates exist, they may be imprecise because of ill-conditioning of the covariance matrix. Moreover, choosing a model that contains a large number of predictors runs the risk of overfitting the data. Future predictions will then tend to be poorer than those obtained with a more parsimonious model.

As in ordinary model selection using ML, variable selection algorithms such as forward selection and backward elimination have pitfalls, especially when p is large. For example, for the set of predictors having no true effect, the maximum sample correlation with the response can be quite large. Also, there can be spurious collinearity among the predictors or spurious correlation between an important predictor and a set of unimportant predictors, because of the high dimensionality[3]. Other criteria exist for identifying an optimal subset of explanatory variables, such as minimizing prediction error or (with AIC) considering models with nearly minimum Kullback–Leibler divergence of the fitted values from true conditional means. With large p, though, it is not feasible to check a high percentage of the possible subsets of predictors, and the danger remains of identifying an effect as important that is actually spurious.

[3]Figure 1 in Fan and Lv (2010) illustrates these issues.

Ordinary variable selection methods such as stepwise procedures are highly discrete: Any particular variable either is or is not selected. Penalized likelihood is more continuous in nature, with some variables perhaps receiving little influence in the resulting prediction equation but not being completely eliminated. Besides providing shrinkage of parameter estimates, some of those methods (L_q-norm with $0 \leq q \leq 1$) also help with variable selection. With the lasso ($q = 1$), many explanatory variables receive zero weight in the prediction equation, the number included depending on the smoothing parameter. A variable can be eliminated, but in a more objective way that does not depend on which variables were previously eliminated.

The variable selection methods for large p fall roughly into two types. One approach adapts dimension-reduction methods, such as stepwise methods and penalized likelihood and regularization using L_q-norm penalties for some q between 0 and 2 and compromise norms. A second approach attempts to identify the relevant effects using standard significance tests but with an adjustment for multiplicity. A fundamental assumption needed for methods to perform well with large p is *sparse structure*, with relatively few elements in $\boldsymbol{\beta}$ being nonzero (Bühlmann et al. 2014).

The first type of variable selection method includes stepwise methods that use regularization procedures. The LARS (least-angle regression) procedure (Efron et al. 2004) for linear models is an adaptation of a forward selection method. Like forward selection, it first adds the predictor having greatest absolute correlation with y, say x_j. This is the variable with the smallest angle between it and the response variable, found for the vectors connecting the origin to the points $\boldsymbol{y}$ and $\boldsymbol{x}_j$ in $\mathbb{R}^n$. The LARS algorithm proceeds from the origin in the x_j direction as long as the angle between the point on that line and the residual between $\boldsymbol{y}$ and that point is smaller than the angle between other predictors and the residual. When some other predictor, say x_k, has as much correlation with the current residual, instead of continuing along the x_j direction, LARS then proceeds in a direction equiangular between x_j and x_k. The algorithm continues in this direction until a third variable x_ℓ earns its way into the "most correlated" set. LARS then proceeds equiangularly between x_j, x_k, and x_ℓ (i.e, along the "least angle direction") until a fourth variable enters, and so on. It smoothly blends in new variables rather than adding them discontinuously.

Advantages of the LARS method are that it is computationally fast and the lasso can be generated in a modified special case. In the published discussion for the Efron et al. article, D. Madigan and G. Ridgeway suggested an extension for logistic regression, and S. Weisberg suggested caution, arguing that any automatic method relying on correlations has potential pitfalls, especially under collinearity. In the rejoinder, the authors discussed possible stopping rules for the algorithm.

An alternative approach that explicitly performs dimension reduction is *principal component analysis*. This method[4] replaces the p predictors by fewer linear combinations of them (the "principal components") that are uncorrelated. The first principal component is the linear combination (using a unit vector) that has the largest possible variance. Each succeeding component has the largest possible variance under the constraint that it is orthogonal to the preceding components. A small number of

[4]Proposed by K. Pearson in 1901 and developed by H. Hotelling in 1933.

principal components often explains a high percentage of the original variability. The components depend on the scaling of the original variables, so when they measure inherently different characteristics they are standardized before beginning the process. Disadvantages, especially with large p, are that it may be difficult to interpret the principal components, and the data may be overfitted, with the derived principal components not explaining variability in another dataset nearly as well. For details, see references in Note 11.3.

The second type of approach searches for effects while adjusting for the number of inferences conducted. This can reduce dramatically the data dimensionality by eliminating the many predictors not having strong evidence of an effect. An approach such as using the *false discovery rate* (FDR) introduced in Section 3.5.3 is especially useful in applications in which a very small proportion of the effects truly are of substantive size. Because of its lessened conservatism and improved power compared with family-wise inference methods such as the Bonferroni, controlling FDR is a sensible strategy to employ in exploratory research involving large-scale testing. A place remains for traditional family-wise inference methods in follow-up validation studies involving the smaller numbers of effects found to be significant in the exploratory studies. Dudoit et al. (2003) surveyed these issues in the context of microarray experiments.

11.2.2 Effect of Large p on Bayesian Methods

Dealing with large p is also challenging for Bayesian inference, perhaps even more so than for frequentist inference. The impact of forming prior distributions for a very large number of parameters may differ from what you intuitively expect. For example, even if you pick a very diffuse prior, the effect may depend strongly on which diffuse prior you choose.

To illustrate, suppose the response distribution is multinomial with p outcome categories and p is very large relative to n, as in a study of the frequency of use of the p words in a language by an author writing in that language. In a particular document, we might observe how many times each word occurs for the n words. Most words would have a count of 0. As in Section 6.1.1, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})$ represent the multinomial trial for observation i, $i = 1, \ldots, n$, where $y_{ij} = 1$ when the response is in category j and $y_{ij} = 0$ otherwise, so $\sum_j y_{ij} = 1$. Let $\pi_{ij} = P(y_{ij} = 1)$, and let $n_j = \sum_i y_{ij}$ denote the total number of observations in category j. Here, for simplicity, we discuss large-p challenges[5] without any reference to explanatory variables, so we will suppress the i subscript and replace π_{ij} by π_j. In practice, similar issues arise when the number of multinomial categories is of any size but the number of explanatory variables is large.

The beta distribution that serves as a conjugate prior distribution for a binomial parameter extends to the *Dirichlet distribution* for multinomial parameters. With hyperparameters $\{\alpha_j\}$, the Dirichlet prior density function is proportional to $\prod_{j=1}^{p} \pi_j^{\alpha_j - 1}$. The posterior density is then also Dirichlet, with parameters $\{n_j + \alpha_j\}$.

[5]Of course, here the actual number of parameters is $p - 1$.

The posterior mean of π_j is $(n_j + \alpha_j)/(n + \sum_k \alpha_k)$. The impact of the prior is essentially to add α_j observations to category j for all j before forming a sample proportion. Most applications use a common value α for $\{\alpha_j\}$, so the impact is to smooth in the direction of the equi-probability model.

The Dirichlet prior with $\alpha = 1$ corresponds to a uniform prior distribution over the probability simplex. This seems diffuse, but it corresponds to adding p observations and then forming sample proportions. This is considerable when p is large relative to n. For example, suppose $n = 100$ but $p = 1000$. The posterior mean of π_j is $(n_j + 1)/(n + p) = (n_j + 1)/1100$. When cell j contains one of the 100 observations, the posterior mean estimate for that cell is 0.0018, shrinking the sample proportion of 0.010 toward the equi-probability value of 0.001. This seems like a reasonable estimate. But what if instead all 100 observations fall in cell j? The posterior mean estimate is then 0.092. This shrinks much more from the sample proportion value of 1.0 than we are likely to believe is sensible. Even though the prior distribution is quite diffuse, it has quite a strong impact on the results. The Jeffreys prior, $\alpha = 1/2$, corresponds to a U-shaped beta density for the binomial case $p = 2$. The shrinkage is then a bit less, but it still gives a posterior mean estimate for π_j of $(n_j + 1/2)/(n + p/2) = (n_j + 1/2)/600$, or 0.1675 when $n_j = n = 100$.

This simplistic example illustrates that the choice of the prior distribution is crucial when p is very large, especially when we depart from the traditional setting in which n is much larger than p. Berger et al. (2013) suggested that the prior distribution should have marginal posterior distributions all close to a common posterior that we'd obtain in the single-parameter case. For instance, we could aim for the posterior distribution of π_j to be approximately a beta distribution with parameters $n_j + \frac{1}{2}$ and $n - n_j + \frac{1}{2}$, which we'd obtain with a Jeffreys prior for the binomial distribution with parameter π_j. We can obtain this by using Dirichlet hyperparameters $\{\alpha_j = 1/p\}$ instead of $\{\alpha_j = 1/2\}$, which is much more diffuse when p is large. This yields a posterior mean for π_j of $(n_j + 1/p)/(n + 1)$. With $n = 100$ observations in $p = 1000$ cells, this is 0.0099 when $n_j = 1$ and is 0.990 when $n_j = 100$.

This approach seems sensible, but even with it, situations exist for which the results may seem inappropriate. When $p = 1000$, suppose we have only $n = 2$ observations, of which $n_j = 1$. The posterior mean for π_j is then 0.334. Would you want to use an estimate that shrinks the sample proportion of 1/2 based on only two observations so little toward the equi-probability value of 0.001? Which prior distribution would you use for such sparse multinomial modeling?

11.3 SMOOTHING, GENERALIZED ADDITIVE MODELS, AND OTHER GLM EXTENSIONS

The models in this text smooth the data rather severely, by producing fitted values satisfying a predictor that is linear in the parameters. In this final section we present frequentist ways of smoothing data that provide more flexibility than linear predictors in GLMs. We also consider alternative models that are nonlinear in the parameters or that describe quantiles instead of mean responses.

11.3.1 Kernel Smoothing

Kernel smoothing, in its basic form, is completely non-model-based. To estimate a mean at a particular point, it smooths the data by using primarily the data at nearby points.

We illustrate with a method that smooths binary response data to portray graphically the form of dependence of y on a quantitative explanatory variable x (Copas 1983). Let $\phi(\cdot)$ denote a symmetric unimodal *kernel function*, such as the standard normal or another bell-shaped pdf. At any x, the kernel-smoothed estimate of $P(y = 1 \mid x)$ is

$$\tilde{\pi}(x) = \frac{\sum_{i=1}^{n} y_i \phi[(x - x_i)/\lambda]}{\sum_{i=1}^{n} \phi[(x - x_i)/\lambda]}, \tag{11.1}$$

where $\lambda > 0$ is a smoothing parameter. At any point x, the estimate $\tilde{\pi}(x)$ is a weighted average of the $\{y_i\}$. For the simple function $\phi(u) = 1$ when $u = 0$ and $\phi(u) = 0$ otherwise, $\tilde{\pi}(x_k)$ simplifies to the sample proportion of successes at $x = x_k$. Then there is no smoothing. When ϕ is proportional to the standard normal pdf, $\phi(u) = \exp(-u^2/2)$, the smoothing approaches this as $\lambda \to 0$. For very small λ, only points near x have much influence. Using mainly very local data produces little bias but high variance. By contrast, as λ increases, data points farther from x also contribute substantially to $\tilde{\pi}(x)$. As λ increases and very distant points receive more weight, the smoothed estimate becomes more like the overall sample proportion. It becomes more highly biased but has smaller variance. As λ grows unboundedly, the smooth function $\tilde{\pi}(x)$ converges to a horizontal line at the level of the overall sample proportion.

For this kernel smoother, the choice of λ is more important in determining $\tilde{\pi}(x)$ than is the choice of ϕ. Copas recommended selecting λ by plotting the resulting function for several values of λ, varying around a value equal to 10 times the average spacing of the x values. The kernel smoothing (11.1) generalizes to incorporate multiple predictors, with a multivariate kernel function such as a multivariate normal pdf.

11.3.2 Nearest-Neighbors Smoothing

In more general contexts than binary regression, smoothers of the kernel type can base estimation at a point on using nearby points. A very simple such method is *nearest-neighbors smoothing*. It is often used for classification, such as by predicting an observation for a subject based on a weighted average of observations for k subjects who have similar values on the explanatory variables.

An advantage of this method is its simplicity, once we select a similarity measure to determine the nearest neighbors. However, the choice of this measure may not be obvious, especially when p is large with possibly some subsets of explanatory variables being highly correlated and some of them being qualitative. More complex smoothers generalize this idea by basing the prediction at a point on a weighted

regression using nearby points, such as described next. Such methods have better statistical properties, such as usually lower bias.

11.3.3 The Generalized Additive Model

The GLM generalizes the ordinary linear model by permitting non-normal distributions and modeling functions of the mean. The quasi-likelihood approach (Chapter 8) generalizes GLMs, specifying how the variance depends on the mean without assuming a particular distribution. Another generalization of the GLM replaces the linear predictor by additive smooth functions of the explanatory variables. The GLM structure $g(\mu_i) = \sum_j \beta_j x_{ij}$ generalizes to

$$g(\mu_i) = \sum_{j=1}^{p} s_j(x_{ij}),$$

where $s_j(\cdot)$ is an unspecified smooth function of predictor j. Like GLMs, this model specifies a link function g and a distribution for y. The resulting model is called a *generalized additive model*, symbolized by GAM (Hastie and Tibshirani 1990). The GLM is the special case in which each s_j is a linear function. Also possible is a mixture of explanatory terms of various types: Some s_j may be smooth functions, others may be linear functions as in GLMs, and others may be indicator variables to include qualitative factors.

A useful smooth function is the *cubic spline*. It has separate cubic polynomials over sets of adjacent intervals for an explanatory variable, joined together smoothly at boundaries of those intervals. The boundary points, called *knots*, can be set at evenly spaced points for each predictor or selected according to a criterion involving both smoothness and closeness of the spline to the data. A *smoothing spline* uses knots at the observed predictor values but imposes a smoothing parameter that determines the influence of the integrated squared second derivative of the smoothing function in penalizing the log likelihood. For example, for the normal model with identity link, the fit minimizes a penalized residual sum of squares,

$$\sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{p} s_j(x_{ij}) \right]^2 + \sum_{j=1}^{p} \lambda_j \int \left[s_j''(x) \right]^2 dx.$$

Larger smoothing parameter values λ_j result in smoother functions (less "wiggling" and change in the first derivative). In fact, this criterion results in a solution in which each s_j is a cubic spline.

One can select λ_j so that a term s_j in the predictor has an *effective df* value, with higher λ_j corresponding to lower effective df. For instance, a smooth function having effective $df = 3$ is similar in overall complexity to a third-degree polynomial, and df close to 1 is similar to a straight line. Choosing an effective df value or a value for a smoothing parameter determines how smooth the resulting GAM fit looks. It is

sensible to try various degrees of smoothing. The goal is not to smooth so much that the fit suppresses interesting patterns yet smooth the data sufficiently so that the data are not overfitted with a highly wiggly function. The smoothing may suggest that a linear model is adequate with a particular link function, or it may suggest ways to improve on linearity.

Using the effective *df* value for each s_j in the additive predictor, we can conduct approximate large-sample inference about those terms. For any model fit, there is a deviance, which reflects the assumed distribution for y. As in comparing GLMs, we can compare deviances for nested GAMs to test whether a particular model gives a significantly better fit than a simpler model.

For fitting a GAM, the *backfitting algorithm* employs a generalization of the Newton–Raphson method that uses local smoothing. The algorithm initializes $\{\hat{s}_j\}$ identically at 0. Then at a particular iteration, it updates the estimate $\hat{s}_j$ by a smoothing of partial residuals $\{y_i - \sum_{k \neq j} \hat{s}_k(x_{ik})\}$ that uses the other estimated smooth functions at that iteration, in turn for $j = 1, \ldots, p$.

An alternative way to smooth the data, without making a distributional assumption for y, employs a type of regression that gives greater weight to nearby observations in predicting the value at a given point; such *locally weighted least squares regression* is often referred to as *lowess* (Cleveland 1979). We prefer GAMs to lowess, because they recognize explicitly the form of the response variable. For instance, with a binary response, lowess can give predicted values below 0 or above 1 at some predictor settings. This cannot happen with a GAM that assumes a binomial random component.

Smoothing methods such as GAMs have the advantage over GLMs of greater flexibility. Using them, we may discover patterns we would miss with ordinary GLMs, and we obtain potentially better predictions of future observations. The smoothness of the function that works well is summarized by its effective df. A disadvantage of GAMs (and other smoothing methods) compared with GLMs is the loss of interpretability for describing the effect of an explanatory variable that has a smooth term in the predictor. Likewise, it is unclear how to apply confidence intervals to effects in a GAM. So it is more difficult to judge when an effect has substantial importance. Thus, when suitable, GLMs are ideal for statistical inference. Also, because any smoothing method has potentially a very large number of parameters, it can require a large n to estimate the functional form accurately.

Even if you plan mainly to use GLMs, a GAM is helpful for exploratory analysis. For instance, for binary responses, scatterplots are not very informative. Plotting the fitted smooth function for a predictor may reveal a general trend without assuming a particular functional relation, as we illustrate in Section 11.3.5.

11.3.4 How Much Smoothing? The Bias–Variance Tradeoff

Smoothing methods have a nonparametric flavor, because they base analyses on a more general structure than a linear predictor. However, in some ways the demands are greater: We need to choose among a potentially infinite number of forms relating the

response variable to the explanatory variables, the number of parameters is potentially much larger, and overfitting is a danger.

As discussed in Section 4.6.2, model selection is at the heart of the fundamental statistical tradeoff between bias and variance. Using a particular model has the disadvantage of increasing the potential bias (e.g., a true mean differing from the value corresponding to fitting the model to the population), but it has the advantage that the parsimonious limitation of the parameter space results in decreased variance in estimating characteristics of interest. The methods presented in this chapter provide a compromise. A method typically starts with a model and its likelihood function, but smooths results to adjust for ways an ordinary linear predictor may fail. All smoothing methods require input from the methodologist to control the degree of smoothness imposed on the data in order to deal with the bias–variance tradeoff, whether it be determined by a smoothing parameter in a frequentist approach or a prior distribution in a Bayesian approach.

11.3.5 Example: Smoothing to Portray Probability of Kyphosis

Hastie and Tibshirani (1990, p. 282) described a study to determine risk factors for kyphosis, which is severe forward flexion of the spine following corrective spinal surgery. Figure 11.3 shows this binary outcome y (1 = kyphosis present, 0 = absent) plotted against the age in months at the time of the operation. For the youngest and the oldest children, most observations have kyphosis absent.

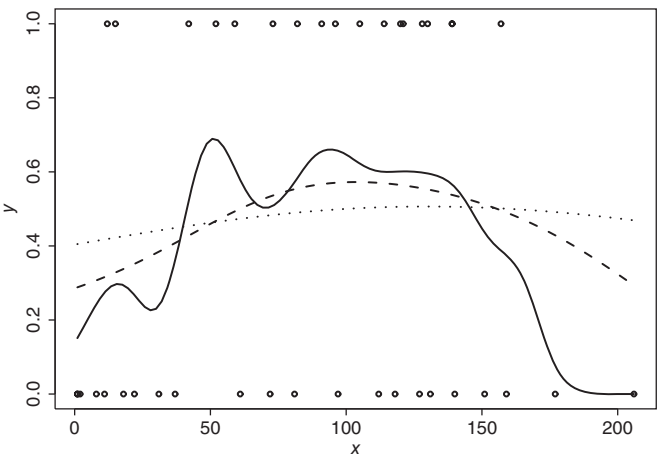

Figure 11.3 Kernel-smoothing estimate of probability of kyphosis as a function of x = age (in months), using standard normal kernel function ϕ and smoothing parameter $\lambda = 25$ (solid curve), 100 (dashed curve), 200 (dotted curve), in equation (11.1).

Figure 11.3 also shows the result of kernel smoothing of the data using the smoother (11.1). The smoothing parameter value $\lambda = 25$ is too low, and the figure

is more irregular than the data justify. The higher values of λ give evidence of nonmonotonicity in the relation. In fact, adding a quadratic term to the standard logistic regression model provides an improved fit.

```
------------------------------------------------------------------------
> Kyphosis # File Kyphosis.dat at www.stat.ufl.edu/~aa/glm/data
     x y
1   12 1
2   15 1
...
40 206 0
> attach(Kyposis)
> plot(x, y)
> k1 <- ksmooth(x, y, "normal", bandwidth=100)
> lines(k1)
> x2 <- x*x
> summary(glm(y ~ x + x2, family=binomial(link=logit)))
              Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -2.046255    0.994348   -2.058    0.0396
x             0.060040    0.026781    2.242    0.0250
x2           -0.000328    0.000156   -2.097    0.0360
---
Residual deviance: 48.228  on 37  degrees of freedom
------------------------------------------------------------------------
```

For fitting a GAM, we treat the data as binomial with a logit link. The default smoothing obtained with the function for GAMs in the VGAM R library falls between a quadratic and cubic in complexity ($df = 2.6$).

```
------------------------------------------------------------------------
> library(VGAM)
> gam.fit <- vgam(y ~ s(x),family=binomialff(link=logit),data=Kyphosis)
> plot(x, fitted(gam.fit))
> summary(gam.fit)
Residual deviance:  47.948 on 35.358 degrees of freedom
DF for Terms and Approximate Chi-squares for Nonparametric Effects
            Df  Npar Df  Npar Chisq  P(Chi)
(Intercept)  1
s(x)         1      2.6      4.7442  0.1528
------------------------------------------------------------------------
```

Figure 11.4 shows the fitted values for the 40 observations. This also suggests using a logistic model with a quadratic term[6]. The figure also shows that fit, which is very similar graphically and in the residual deviance.

We can also fit GAMs using the gam and mgcv libraries in R. We next fit models that have successively linear, quadratic, and cubic complexity for the smooth function.

[6]Using a penalized-likelihood approach for GAMs, Eilers and Marx (2002) suggested instead a bell-shaped response curve.

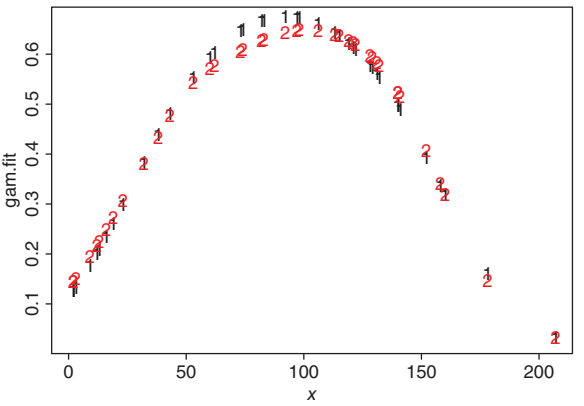

Figure 11.4 Estimates of probability of kyphosis as a function of x = age, using (1) a GAM and (2) logistic regression with quadratic term.

Comparison of deviances shows that quadratic fits better than linear, but cubic is not better than quadratic.

```
-------------------------------------------------------------------------
> library(gam) # R library by Trevor Hastie
> gam.fit1 <- gam(y ~ s(x,1), family=binomial, data=Kyphosis)
> gam.fit2 <- gam(y ~ s(x,2), family=binomial, data=Kyphosis)
> gam.fit3 <- gam(y ~ s(x,3), family=binomial, data=Kyphosis)
> anova(gam.fit1, gam.fit2, gam.fit3)
Analysis of Deviance Table
Model 1: y ~ s(x, 1) # linear complexity
Model 2: y ~ s(x, 2) # quadratic
Model 3: y ~ s(x, 3) # cubic
  Resid.  Df  Resid. Dev       Df  Deviance  Pr(>Chi)
1         38      54.504
2         37      49.216   0.9999    5.2880    0.0215
3         36      48.231   1.0002    0.9852    0.3210
-------------------------------------------------------------------------
```

11.3.6 Quantile Regression

Models in this book describe the conditional mean of y as a function of explanatory variables. Alternatively, we could model a quantile. For example, in modeling growth over time for a sample of a biological organism, it might be of interest to estimate the 10th percentile, median, and 90th percentile of the conditional distribution as a function of time. *Quantile regression* models quantiles of a response variable as a function of explanatory variables. M-estimation in regression using $\rho(e_i) = |e_i|$ corresponds to quantile regression for the median.

Like regression fitted by M-estimation, this method can be less severely affected by outliers than is ordinary least squares. When the response conditional distributions are highly skewed with possibly highly nonconstant variance, the method can describe the

relationship better than a simple normal model with constant variance. For instance, consider modeling of annual income as a function of the age of a person. We might expect almost no effect at low quantiles, with the effect increasing as the quantile increases, reflecting also increasing variability with age.

Quantile-regression model fitting minimizes a weighted sum of absolute residuals, formulated as a linear programming problem. This is available in software[7]. Why not always use it instead of least squares, since it is less affected by outliers? When the normal linear model truly holds, the least squares estimators are much more efficient.

11.3.7 Nonlinear Regression

In this book, we've focused on predictors that are linear in the parameters. The GAM is one generalization. Another is relevant for applications in which the predictor is naturally nonlinear in parameters. For example, consider the model

$$E(y_i) = \frac{\beta_0}{1 + \exp[-(\beta_1 + \beta_2 x_i)]}.$$

With $\beta_0 = 1$, this is the logistic regression curve for a binary response probability (Chapter 5). For other β_0, it has a symmetric, sigmoidal shape with bounds of 0 and β_0. This model can describe the growth of a tumor or population growth, when the maximum possible size is also a parameter.

A *nonlinear regression model* has the form

$$E(y_i) = f(\boldsymbol{x}_i; \boldsymbol{\beta}),$$

where f is a known function of the explanatory variables and the parameters. With an assumption about the distribution of y_i, inference can use likelihood-based methods. Assuming normality with constant variance σ^2, this again yields the least squares criterion, with $\hat{\boldsymbol{\beta}}$ giving the minimum value of $\sum_i [y_i - f(\boldsymbol{x}_i; \boldsymbol{\beta})]^2$. The likelihood equations are then

$$\sum_{i=1}^{n} [y_i - f(\boldsymbol{x}_i; \boldsymbol{\beta})] \frac{\partial f(\boldsymbol{x}_i; \boldsymbol{\beta})}{\partial \beta_j} = 0, \quad j = 1, \ldots, p.$$

Finding the estimates requires an iterative algorithm that starts at initial values $\boldsymbol{\beta}^{(0)}$ for $\hat{\boldsymbol{\beta}}$. Let $\boldsymbol{X}$ denote the model matrix of $\{x_{ij}\}$, and let $\boldsymbol{f}(\boldsymbol{X}; \boldsymbol{\beta})$ be the vector having elements $f(\boldsymbol{x}_i; \boldsymbol{\beta})$. The *Gauss–Newton algorithm*, which for a normal response is equivalent to Fisher scoring, uses the linearization

$$\boldsymbol{f}(\boldsymbol{X}; \boldsymbol{\beta}) = \boldsymbol{f}(\boldsymbol{X}; \boldsymbol{\beta}^{(0)}) + \boldsymbol{G}^{(0)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}),$$

[7] Such as the `quantreg` package in R.

where the gradient matrix $\boldsymbol{G}^{(0)}$ has elements $\partial f(\boldsymbol{x}_i;\boldsymbol{\beta})/\partial\beta_j$ evaluated at $\boldsymbol{\beta}^{(0)}$. For the initial working residuals $\{\boldsymbol{e}_i^{(0)} = y_i - f(\boldsymbol{x}_i;\boldsymbol{\beta}^{(0)})\}$, the first iteration yields updated estimate

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + (\boldsymbol{G}^{(0)\mathrm{T}}\boldsymbol{G}^{(0)})^{-1}\boldsymbol{G}^{(0)\mathrm{T}}\boldsymbol{e}^{(0)}.$$

Each subsequent iteration t regresses the current working residuals $\boldsymbol{e}^{(t)}$ on the current gradient matrix $\boldsymbol{G}^{(t)}$ to find the increment to the working estimate $\boldsymbol{\beta}^{(t)}$. Modifications of the method exist, such as taking smaller increments if needed to decrease the residual sums of squares, or using numerical derivatives rather than computing the gradient matrix. Many nonlinear models have the potential for multiple local maxima of the log likelihood, so it is wise to use a grid of quite different initial values to increase the chance of finding the true least squares estimate.

The linearization is also the basis of standard errors for $\hat{\boldsymbol{\beta}}$. The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\right)^{-1},$$

where the gradient matrix $\boldsymbol{G}$ has elements $\partial f(\boldsymbol{x}_i;\boldsymbol{\beta})/\partial\beta_j$ evaluated at $\boldsymbol{\beta}$. This has the same form as $\mathrm{var}(\hat{\boldsymbol{\beta}})$ in formula (2.4) for the ordinary linear model, except that the gradient matrix replaces the model matrix. In practice, we estimate the covariance matrix by substituting $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ in $\boldsymbol{G}$ and by estimating σ^2 by the error mean square $s^2 = \sum_i[y_i - f(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}})]^2/(n-p)$. Nonlinear regression fitting methods and subsequent inference are available in software[8].

CHAPTER NOTES

Section 11.1: Robust Regression and Regularization Methods for Fitting Models

11.1 **Robust regression**: M-estimation evolved out of research by Huber (1964) on robust estimation of a location parameter. Huber and Ronchetti (2009, Chapter 7) presented the regression context of this approach. Rousseeuw (1984) proposed another alternative to least squares, finding the estimate that produces the smallest *median* of the squared residuals, instead of their *mean* (equivalently, sum). Like M-estimation with $\rho(e_i) = |e_i|$, this "least median of squares" method can have low efficiency when the ordinary normal linear model nearly holds. Cantoni and Ronchetti (2001) proposed other robust estimation methods for GLMs. Birkes and Dodge (1993) surveyed alternative fitting methods, including least-absolute-deviations regression, M-estimation, and ridge regression.

11.2 **Ridge regression and lasso**: For more on ridge regression, see Hoerl and Kennard (1970) and Hastie et al. (2009, Section 3.4.1). For the lasso, see Bühlmann and van de Geer (2011, Chapters 2 and 3), Hastie et al. (2009, Section 3.4.2), Izenman (2008), and Tibshirani (1996, and his website `statweb.stanford.edu/~tibs/lasso.html`). James et al. (2013) gave a less technical introduction to such methods,

[8] such as the `nls` function in R.

with extensive R examples. The bootstrap is another possible way to determine standard errors for lasso estimates (Chatterjee and Lahiri 2011). Lockhart et al. (2014) proposed a significance test for the lasso, based on how much of the covariance between y and the model-fitted values can be attributed to a particular predictor when it enters the model. Bühlmann et al. (2014) presented other inference methods, such as tests based on multisample splitting of the data.

Section 11.2: Modeling with Large p

11.3 **Penalized likelihood with large** p: Fan and Lv (2010) and Tutz (2011) reviewed penalized likelihood methods for variable selection in high dimensions. Fan and Lv noted that the lasso has a tendency to include many false-positive variables when p is large. For details about dimension-reduction methods such as principal component analysis, see Hastie et al. (2009, Chapter 18), Izenman (2008), and James et al. (2013). Bühlmann et al. (2014) presented a brief introductory survey of high-dimensional methods. For multinomial modeling, see Taddy (2013).

11.4 **Bayes with large** p: For issues in selecting priors when p is large but n may not be, see Berger et al. (2013), Griffin and Brown (2013), Kass and Wasserman (1996, Section 4.2.2), and Polson and Scott (2010). Carvalho et al. (2010) advocated a prior based on multivariate normal scale mixtures. Gelman (2006) argued for using noninformative priors (such as uniform) for variance parameters in hierarchical models. For variable selection issues, see George and McCulloch (1997), George (2000), and Rŏcková and George (2014). Hjort et al. (2010) surveyed nonparametric Bayesian approaches.

Section 11.3: Smoothing, Generalized Additive Models, and Other GLM Extensions

11.5 **Smoothing**: For smoothing methods, see Fahrmeir et al. (2013, Chapter 8), Fahrmeir and Tutz (2001, Chapter 5), and Faraway (2006, Chapter 11). Green and Silverman (1993), Hastie et al. (2009), Izenman (2008), James et al. (2013), Simonoff (1996), Tutz (2011, Chapters 6 and 10), and Wakefield (2013, Chapters 10–12). For smoothing spatial data , see Fahrmeir et al. (2013, Section 8.2). Albert (2010) presented Bayesian smoothing methods.

11.6 **GAMs and penalized-spline regularization**: For generalized additive modeling, see Fahrmeir et al. (2013, Chapter 9), Faraway (2006, Chapter 12), Hastie and Tibshirani (1990), Wood (2006), and Yee and Wild (1996). The *generalized additive mixed model* adds random effects to a GAM (Wood 2006, Chapter 6). Eilers and Marx (1996, 2002, 2010) introduced an alternative penalized likelihood approach for splines that provides a way of fitting GAMs as well as a mechanism for regularization. Rather than penalizing by the integrated squared derivative, it penalizes by differences of coefficients of adjacent splines. See also Fahrmeir et al. (2013, Chapter 8).

11.7 **Nonlinear and quantile regression**: For nonlinear regression methods, see Bates and Watts (1988) and Seber and Wild (1989). For a brief review, see Smyth (2002). For quantile regression and examples, see Davino et al. (2013), Fahrmeir et al. (2013, Chapter 10), and Koenker (2005).

11.8 **Functions and images**: Methods of this chapter extend to the analysis of more complex types of data, such as functions and images. See, for example, Crainiceanu et al. (2009), Ramsay and Silverman (2005), Di et al. (2009), and `www.smart-stats.org` and the R package `refund`.

EXERCISES

11.1 Show that M-estimation with $\rho(e_i) = |e_i|$ gives the ML solution assuming a Laplace distribution for the response.

11.2 The *breakdown point* of an estimator is the proportion of observations that must be moved toward infinity in order for the estimator to also become infinite. The higher the breakdown point, the more robust the estimator. For estimating the center of a symmetric distribution, explain why the breakdown point is $1/n$ for the sample mean but 0.50 for the sample median. (However, even robust regression methods, such as using $\rho(e_i) = |e_i|$, can have small breakdown points or other unsatisfactory behavior; see Seber and Lee 2003, Sections 3.13.2 and 3.13.3)

11.3 Refer to the equations solved to obtain $\hat{\boldsymbol{\beta}}$ for M-estimation and the expression for var$(\hat{\boldsymbol{\beta}})$. Show how they simplify for least squares.

11.4 In M-estimation, let $\rho(x) = 2(\sqrt{1 + x^2/2} - 1)$. Find the influence function, and explain why this gives a compromise between least squares and $\rho(x) = |x|$, having a bounded influence function with a smooth derivative at 0.

11.5 Since the Gauss-Markov theorem says that least squares estimates are "best," are not estimates obtained using M-estimation necessarily poorer? Explain.

11.6 For the saturated model, $E(y_i) = \beta_i$, $i = 1, \ldots, n$, find the ridge-regression estimate of β_i and interpret the impact of λ.

11.7 For the normal linear model, explain how (**a**) the ridge-regression estimates relate to Bayesian posterior means when $\{\beta_j\}$ have independent $N(0, \sigma^2)$ distributions, (**b**) the lasso estimates relate to Bayesian posterior modes when $\{\beta_j\}$ have independent Laplace (double-exponential) distributions with means 0 and common scale parameter.

11.8 Consider the linear model

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{40} x_{i,40} + \epsilon_i$$

with $\beta_1 = 1$ and $\beta_2 = \cdots = \beta_{40} = 0$, where $x_{ij} = u_i + v_j$ with $\{u_i\}$, $\{v_j\}$, and $\{\epsilon_i\}$ being iid $N(0, 1)$ random variables.

a. Find the correlation between y and x_1, y and x_j for $j \neq 1$, and x_j and x_k for $j \neq k$, and the multiple correlation between y and the set of explanatory variables.

b. Using this model, randomly generate $n = 100$ observations on the 41 variables. Use the lasso to select a model, for a variety of λ smoothing

parameter values. Summarize results, and evaluate the effectiveness of this method.

c. Specify alternative values for $\{\beta_j\}$ for which you would not expect the lasso to be effective. Re-generate $\boldsymbol{y}$, and summarize results of using the lasso.

11.9 Refer to the Dirichlet prior distribution introduced for multinomial parameters in Section 11.2.2. Explain why a multivariate normal prior for multinomial logits provides greater flexibility.

11.10 Refer to Copas's kernel smoother (11.1) for binary regression, with $\phi(u) = \exp(-u^2/2)$.

a. To describe how close this estimator falls at a particular x value to a corresponding smoothing in the population, use the delta method to show that an estimated asymptotic variance is

$$\tilde{\pi}(x)[1-\tilde{\pi}(x)]\frac{\sum_i \phi\left[\sqrt{2}(x-x_i)/\lambda\right]}{\left\{\sum_i \phi[(x-x_i)/\lambda]\right\}^2}.$$

Explain why this decreases as λ increases, and explain the implication.

b. As λ increases unboundedly, explain intuitively to what $\tilde{\pi}(x)$ and this estimated asymptotic variance converge.

11.11 When $p > n$, why is backward elimination not a potential method for selecting a subset of explanatory variables?

11.12 Sometimes nonlinear regression models can be converted to ordinary GLMs by employing a link function for the mean response and/or transforming the explanatory variables. Explain how this could be done for the normal-response models (**a**) $E(y_i) = \beta_0 \exp(\beta_1 x_i)$ (an exponential growth model), (**b**) $E(y_i) = 1/(\beta_0 + \beta_1 x_i + \beta_2 x_i^2)$.

11.13 Refer to the form of iterations for the Gauss–Newton algorithm described in Section 11.3.7. Show that an analogous formula

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\epsilon}$$

holds for the ordinary linear model, where $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{\mu}$ is a "true residual."

11.14 Randomly generate nine observations satisfying a normal linear model by taking $x_i \sim N(50, 20)$ and $y_i = 45.0 + 0.1x_i + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$. Now add to the dataset a contaminated outlying observation 10 having $x_{10} = 100$, $y_{10} = 100$. Fit the normal linear model to the 10 observations using (**a**) least

squares, (**b**) Huber's M-estimation. Compare the model parameter estimates and the estimate of σ. Interpret.

11.15 For the data analyzed in Section 11.1.8 on risk factors for endometrial cancer, compare the results shown there with those you obtain using the lasso.

11.16 For the horseshoe crab data introduced in Section 1.5.1 and modeled in Section 7.5, suppose you use graphics to investigate how a female crab's number of male satellites depends on the width (in centimeters) of the carapace shell of the crab. If you plot the response counts of satellites against width, the substantial variability makes it difficult to discern a clear trend. To get a clearer picture, fit a generalized additive model, assuming a Poisson distribution and using the log link. What does this suggest about potentially good predictors and link functions for a GLM?

11.17 For the horseshoe crab dataset, Exercise 5.32 used logistic regression to model the probability that a female crab has at least one male satellite. Plot these binary response data against the crab's carapace width. Also plot a curve based on smoothing the data using a kernel smoother or a generalized additive model, assuming a binomial response and logit link. (This curve shows a roughly increasing trend and is more informative than viewing the binary data alone.)

11.18 For the `Housing.dat` file analyzed in Sections 3.4 and 4.7, use methods of this chapter to describe how the house selling price depends on its size.

11.19 Continue the previous exercise, now using all the explanatory variables.