Statistics Surveys
Vol. 6 (2012) 1–73
ISSN: 1935-7516
DOI: 10.1214/12-SS100

Prediction in several conventional contexts*

Bertrand Clarke[†,‡]

Department of Medicine
University of Miami
1120 NW 14 Street, Suite 611 (CRB C-213)
Miami, FL, 33136, USA
e-mail: bclarke2@med.miami.edu

and

Jennifer Clarke[†]

Department of Epidemiology and Public Health
University of Miami
1120 NW 14th Street Suite 1051 (CRB R-26)
Miami, FL, 33136, USA
e-mail: jclarke@biostat.med.miami.edu

Abstract: We review predictive techniques from several traditional branches of statistics. Starting with prediction based on the normal model and on the empirical distribution function, we proceed to techniques for various forms of regression and classification. Then, we turn to time series, longitudinal data, and survival analysis. Our focus throughout is on the mechanics of prediction more than on the properties of predictors.

AMS 2000 subject classifications: Primary 62M20.
Keywords and phrases: Prediction, prequential, IID data, time series, longitudinal data, survival analysis.

Received March 2012.

Contents

*This paper was accepted by Subhashis Ghosal.

†Center for Computational Science, U. of Miami, 1120 NW 14 Street, Suite 610G (CRV C-213) Miami, FL, 33136, USA.

‡Department of Epidemiology and Public Health, U. of Miami. Miami, FL, 33136, USA.

1. Introduction

The stance of this paper is predictive and operational. That is, *we take prediction as our fundamental task and present a large number of ways to accomplish it*. This means that we will focus on how to form predictors in various settings rather than the properties of those predictors let alone discussing estimators, tests, or other statistical objects. For instance, in Sec. 2.1 we will recast the problem of estimating a mean into identifying point predictors for future outcomes. Instead of being concerned with standard errors or posterior variances, we will focus on identifying prediction intervals for future outcomes. Instead of considering classes of estimators or tests, we will consider classes of predictors. Point predictors and prediction intervals are clearly related to estimators and their properties; however, these relationships will not be the focus here. The role

of parameters in our context will mostly be to quantify how well a predictor performs. We argue an overall approach to statistics can be based on prediction and the main point of this paper is to demonstrate this.

Prediction is important for several reasons. First and foremost, in many settings we really are concerned primarily with prediction, not with estimating or testing, even if answers are phrased that way. For instance, one may estimate a probability of recurrence of cancer (with an standard error), but it would be more informative to give a point predictor for when a patient will get a recurrence along with an assessment of the variability of that prediction. If nothing else, the prediction interval is less abstract and more intelligible for most people than a confidence or credible interval.

Second, using prediction based approaches means our inferences are testable and hence any theories they may represent are testable. Testability is not the same as interpretability, but a good predictor will typically permit some, perhaps limited, interpretation. For instance, given a predictor that uses explanatory variables, see Sec. 2.2, one can often determine which of the explanatory variables are most important for good prediction. More generally, apart from interpretability, theories for physical phenomena that arise from estimating a model and using hypothesis tests to simplify it must be validated predictively. A propos of this, a criticism of predictive approaches used to justify direct modeling approaches is that being able to predict well does not imply that the phenomenon in question is understood. The answer to this criticism is that modeling only implies understanding when the model has been extensively validated i.e., found to be true, and this validation is primarily predictive. So, announcing a model before doing extensive validation – as is typically done – only provides the illusion of understanding. Prediction is a step toward model building, not the reverse, and predictive evaluation is therefore more honest.

Third, prediction in and of itself does not require an unseen world of abstract population quantities or measure spaces. Predictors such as "tomorrow's average temperature will be the same as today's average temperature" (which is probably not a bad predictor) do not require anything we have not measured. We may wish to invoke the rigor of measure theory to provide a theoretical evaluation of our prediction methods under various assumptions but this is a separate task from prediction per se. Indeed, in many cases the asymptotic properties of predictors, in terms of sample size or other indices, are of interest but cannot be obtained without making assumptions that are hard to verify in reality.

In the rest of this section, we define the sequential prediction setting, give examples of classes of predictors, and briefly discuss some of the properties of prediction in general, focusing on the Prequential Principle. At the end of this section, we describe the contents of the rest of the paper.

1.1. Definition and examples of predictor types

To be precise, we now define the paradigmatic sequential prediction problem. The term *prediction* will be reserved for identifying outcomes of random variables while *estimation* is reserved for identifying non-random quantities. It is

not a requirement that the random variable being predicted be observable; in time series it is important to examine residuals and in mixed models examining unmeasured individual deviations from a population mean is the motivation for the model. Suppose we want to predict a sequence of random variables $Y_1, \ldots, Y_n, \ldots$. To make the problem more useful, we often try to model outcomes of Y as functions of explanatory variables (covariates) $x = (x_1, \ldots, x_d)$. Thus we define a prediction for the $n+1$ stage by

$$\hat{Y}_{n+1} = \hat{Y}_{n+1}(x_{n+1}) \tag{1.1}$$

where we assume the values $x_{n+1} = (x_{1,n+1}, \ldots, x_{d,n+1})$ are available before the prediction $\hat{Y}_{n+1}$ must be made. We also assume that all previous data (x_i, Y_i) for $i = 1, \ldots, n$ are available to help in the construction of $\hat{Y}_{n+1}$. Without loss, here n may be regarded as (discrete) time, but more generally n can be any index that orders the sequence of predictions to be made. While n may also be continuous we do not consider that possibility. Note that (1.1) is phrased as one step ahead prediction, but if the goal is to predict q steps ahead, the formulation is similar: We require a vector valued function $\hat{Y}_{n+1,\ldots,n+q} = (\hat{Y}_{n+1}, \ldots, \hat{Y}_{n+q})$ and the individual functions $\hat{Y}_{n+j}$ for $j = 1, \ldots, q$ are again constructed using all the data (x_i, Y_i) for $i = 1, \ldots, n$. We regard the sequential prediction context as foundational because it permits the predictor to evolve as data accumulate and threfore provides a more demanding evaluation of any prediction strategy, i.e., choice of the sequence $\hat{Y}_{i+1}$ for $i = 2, 3, \ldots$.

Estimation and prediction provide inference about different quantities and the analysis of a problem depends on whether one takes a predictive perspective or a parameter inference perspective. An example may help to clarify this. Suppose we have independent measurements on one subject, say $y_1, \ldots, y_k$. We can use $\bar{y}$ to estimate $\mu_Y = E(Y)$ or we can use $\bar{y}$ to predict Y_{k+1}. As an estimator, $\bar{Y} \to \mu_Y$ in probability (for instance) but as a predictor, $\bar{Y} - Y_{k+1}$ converges in distribution as $k \to \infty$ to the distribution of $\mu_Y - Y$, where Y is an independent copy of any of the Y's.

As an extension of this, consider a hierarchical experiment where we choose a subject θ from a continuous population according to distribution $w(\theta)$ and again make k IID measurements $y_{\theta 1}, \ldots, y_{\theta k}$ of Y_θ. Now, $\bar{y}_\theta$ estimates $E(Y_\theta)$ and could be used as a predictor for an independent copy of Y_θ (holding θ fixed). However, the natural quantity to estimate would be $\mu = \int w(\theta) E(Y_\theta) d\theta$ and the natural random variable to predict would be $Y \sim \int w(\theta) p(\cdot|\theta) d\theta$. The obvious estimator of μ would no longer be any of the individual $\bar{y}_\theta$'s but rather $(1/m)\sum_{i=1}^m \bar{y}_{\theta_i}$, the grand mean over the m subjects sampled. For prediction, we get $(1/m)\sum_{i=1}^m \bar{y}_{\theta_i} - Y$ converges to $\mu - Y$ in distribution as $m, k \to \infty$. That is, the population generating the random variable to be predicted has changed from measurements on an individual to members of the population from which the individual was drawn with consequent changes for estimation and prediction.

The simplest class of predictors arises when there are no explanatory variables and Y follows a parametric family. For instance, given a parametric fam-

ily of densities $p(\cdot|\theta)$ for a random variable Y, with respect to a dominating measure, and indexed by a d-dimensional parameter $\theta \in \mathbb{R}^d$, one can collect an independent and identical (IID) sample of size n $Y_1 = y_1, \ldots, Y_n = y_n$. The data can be used to give a value for an estimate $\hat{\theta} = \hat{\theta}(y^n)$ of θ where $y^n = (y_1, \ldots, y_n)$ is a realized value of $Y^n = (Y_1, \ldots, Y_n)$ i.e., $Y^n = y^n$. Then one can make predictions for Y_{n+1} using $p(y_{n+1}|\hat{\theta})$. These are called plug-in predictors and it is seen that every distinct estimator leads to a distinct predictor since a likelihood-based prediction interval with confidence α for Y_{n+1} can be obtained from $\{y|\ p(y|\hat{\theta}) > t_\alpha\}$ where t_α is a threshold to give $1-\alpha$ conditional confidence (given $Y^n = y^n$). If predictions are made using $p(y_{n+1}|\hat{\theta})$, they will have variability due to $\hat{\theta}$ as well as due to the intrinsic variability of Y_{n+1} but the variability due to $\hat{\theta}$ usually goes to zero as $n \to \infty$.

In this parametric setting, there are also predictors that can be identified by using the parametric family and these do not correspond to any plug-in predictor. One approach is to base an interval prediction on the predictive density $m(y_{n+1}|y^n)$ where $w(\cdot)$ is a prior density for θ and

$$m(y^n) = \int w(\theta)p(y^n|\theta)d\theta; \tag{1.2}$$

see [2], is the mixture of distributions often called the marginal for the data. Now, one can use (1.2) to give the point predictor $\hat{Y}_{n+1} = E_{m(\cdot|y^n)}(Y_{n+1})$, where the subscript on E indicates the distribution in which the expectation is taken. Using this $\hat{Y}_{n+1}$ would essentially never be the same as using $p(\cdot|\theta)$ for any θ. More generally one can use the same procedure on a member of the class

$$m_q(y_{n+1}|y^n) = \frac{m_q(y^n, y_{n+1})}{\int m_q(y^n, y_{n+1})dy_{n+1}} \tag{1.3}$$

where

$$m_q(y^n, y_{n+1}) = \left(\int w(\theta)p(y^{n+1}|\theta)^q d\theta\right)^{1/q}.$$

Here, q parametrizes a class of densities and controls how much weight the modes of the densities receive relative to the tails.

A frequentist approximation to $m(y_{n+1}|y^n)$ is

$$\hat{m}(y_{n+1}; y^n) = \sup_\theta\ p(y_{n+1}|\theta)p(y^n|\theta);$$

see [67] that uses this to derive prediction intervals (PI's) for sequences of random variables that follow IID normal, Binomial, Poisson, and exponential distributions and [58] who derives analogous PI's when a sequence of random variables follows a censored Weibull distribution. See also the density predictor studied in [84] obtained from $p(y|\hat{\theta})/c$ where c is the integral of $p(y|\hat{\theta}(y^n))$ over y^n and the conditional predictive likelihood approach of [16] (which can be used for imputation as well as prediction). Loosely, when a parametric family is believed

to contain the true density, the class of plug-in predictors, which is equivalent to the number of distinct estimators of θ, is much smaller than the class of all predictors that might be considered.

In addition to plug-in predictors and other density-based predictors, one standard way to construct predictors is via decision theory. Suppose L is a loss function, i.e., a non-negative function assigning a real number to the discrepancy between an action say $a \in \mathcal{A}$, where $\mathcal{A}$ is the action space, and a 'correct' value. If the goal is to predict Y_{n+1} having seen $Y_1, \ldots, Y_n$ then we might choose $\mathcal{A}$ to be the collection of real numbers (dependent on y^n) and condition on y^n. Then we might find the L-optimal predictive action

$$a^*(y^n) = \arg\min_a \int L(a, y_{n+1}) p(y_{n+1}|y^n) dy_{n+1} \tag{1.4}$$

and use it as a predictor for Y_{n+1}, i.e., set $\hat{Y}_{n+1}(y^n) = a^*(y^n)$. Familiar examples of this include $L(u,v) = (u-v)^2$ which gives the predictor $\hat{Y}_{n+1} = E(Y_{n+1}|Y^n = y^n)$ and $L(u,v) = |u-v|$ which gives $\hat{Y}_{n+1} = \mathsf{med}\ Y_{n+1}$ where med is the median taken in the conditional distribution $p(y_{n+1}|y^n)$. One can regard predictors such as these as density based since they arise from using $p(y_{n+1}|y^n)$. However, the introduction of a loss function, a different class of mathematical object from a density, is a very strong assumption – probably stronger than invoking a prior – and the resulting predictors depend heavily on L. Moreover, the optimality properties that arise in decision theory also depend delicately on the parametric family. This means that the predictors may be non-robust to changes in the parametric family or loss function. Typically, decision-theoretic predictors are only optimal when uncertainty in both L and $p(\cdot|\theta)$ is absent; it is not clear how fast the performance of decision-theoretic predictors deteriorates as uncertainty in L and $p(\cdot|\theta)$ increases.

Obtaining a PI from $p(y_{n+1}|y^n)$ will have $1-\alpha$ confidence in the conditional probability $P(\cdot|Y^n = y^n)$, not in the marginal probability $P_{Y_{n+1}}$ for Y_{n+1}. This may be preferred because the PI's from $p(y_{n+1}|y^n)$ will usually be narrower due to the conditioning on y^n. On the other hand, a different outcome $(y^n)'$ would give a different interval that might be as representative of Y_{n+1} as the one given by y^n. So, the optimality properties of a^* from (1.4) in terms of $p(\cdot|Y^n = y^n)$ or the prediction intervals that a^* would generate by using $p(y_{n+1}|y^n)$ might be less desireable than using the distribution $P_{n+1}(\cdot)$ for Y^{n+1} to get a point prediction and PI for Y_{n+1} because this approach includes the variability of Y^n as well. This will tend to enlarge the PI's but may make them more representative, especially for small sample sizes.

To conclude this subsection, note that if one accepts the premise that most real-world phenomena are more complex than the simple models based on limited data that we tend to use, then a common mistake is underestimating model uncertainty and model mis-specification. So, decision theory, which typically requires a fully specified model and loss, will not usually be good a representation for real phenomena. That is, decision-theory usually is only a convenient simplification – perhaps over-simplifiction – that we invoke to construct a predictor.

Hence, decision-theoretic optimality will not, in general, ensure good prediction. In these settings, the only way to ensure a predictor is good is to validate it on new data.

1.2. Criteria for prediction

Having seen plug-in, density based, and decision-theoretic predictors, it is apparent that there are many classes of predictors and they can be quite large. So, it is helpful to impose criteria on predictors as a way to search a given predictor class for members that are likely to perform well.

In the absence of model uncertainty and model mis-specification it may be reasonable to use model-based methods of evaluation. This is common in decision theory where the risk, and variants on it, are the main criteria to be optimized. For instance, the posterior mean is the Bayes-optimal predictor under squared error loss and has good asymptotic properties when some member of the parametric model class is true because the model is used to define the optimality criterion. Criteria such as unbiasedness, minimum variance, coverage probability, and Type I and II error probabilities, among others, are similar in that they too are phrased in terms of the model and rely on the absence of model uncertainty or mis-specification to be valid.

However, in linear regression problems with model uncertainty it is possible to get smaller prediction errors using a method that is not Bayes-optimal, see [96] and this is mainly seen by looking at the sum of squared 'predictuals' of the form $\hat{Y}_i(x_i) - y_i(x_i)$. It is also quite easy to show examples where Frequentist hypothesis testing is misleading when none of the models under consideration are true, i.e., when model mis-specification is present. In this case, too, incorrect conclusions can be detected when predictions from the inferred model are found to be poor. In general, any time the criterion by which a predictor (or other inferential procedure) is judged depends on the validity of the model class, the predictor (or other inferential procedure) may break down in the presence of model uncertainty or mis-specification. The breakdown can be detected by using the predictor on new data and finding an elevated error.

Thus, one way to ensure that model uncertainty and mis-specification have been included in a predictive approach is to use sequential prediction and invoke the Prequential Principle, see [36]; for a more mathematical treatment see [35] and for a more recent elaboration of key principles that might be appropriate for prediction see [25]. One way the (weak) Prequential Principle can be stated is 'the method of construction of a predictor should be disjoint from the method of evaluation of the predictor's performance'. Thus, if a predictor is generated using a model, the model should not be used in the evaluation of the predictor. Ideally, the worth of a predictor sequence in a sequential prediction setting should be evaluated primarily by looking at the cumulative predictive error, a quantity that merely compares the predictions with the actual data values. Note that whether a predictor is derived under a Bayesian, Frequentist or other inferential paradigm is not important: All that matters is how good a predictor is at predicting.

Imposing the Prequential Principle is a way to prevent dubious assumptions about the true model from unduly influencing how we find predictors and use them for inference. Why is this important? We could disregard the Prequential Principle and use $E_M(\hat{Y} - Y)^2$, for example, to evaluate the performance of a predictor $\hat{Y}$ assuming M to be the true model. However, if there is model mis-specification then the assumed model will be wrong. Sometimes the degree of error is small but it can be large and it is frequently difficult to tell how far an assumed model is from the unknown true model. Thus, a predictor from any given model will routinely be suboptimal (for large enough sample sizes) relative to the true model and the degree of suboptimality will be impossible to assess. Such a predictor may not be very suboptimal if the given model is a simplification of a complex true model and the sample size is small, but in this case our predictor will perform poorly as sample size increases. In addition, by assuming M to be the true model we ignore the uncertainty in M. Finally, because M is unknown we want our estimator to perform well for all models in a neighborhood of the true model that includes some simple models (relative to sample size) and not just under the assumed true model. In short, evaluating the performance of a predictor under a single assumed true model is not in general a real test that can be substituted for validation on new data.

While sequential prediction, i.e., issuing a series of predictors $\hat{Y}_{n+1}, \hat{Y}_{n+2}, \ldots$, each refining the previous predictors (since the sample size is increasing) is obviously useful in contexts where data arrive sequentially, here we think of the sequential prediction context as providing enhanced validation. That is, once a predictor sequence performs poorly enough, we can reformulate the basis on which the predictors are generated. Even for data that does not arrive sequentially the thinking involved in successive refinement under the Prequential Principle is useful as a way to see how the predictive structure evolves as data accumulate and as a sort of internal (to the data) check that the final predictor will be reasonable. Alternatively, the sequential properties of a predictive strategy can be evaluated over several random re-orderings of a batch of data and pooled as an internal validation. We suggest that optimal, or at least good, performance under sequential prediction is a necessary requirement for any inferential strategy.

Conceptually, the Prequential Principle is central to prediction since it is the 'pure' case of comparing predictions with outcomes. So, it is important to bear the Prequential Principle in mind when using the techniques presented in the rest of this article. In practice, there are serviceable approximations to this pure case, such as various versions of cross-validation, which may be more convenient for obtaining information about model fit and generalization error, especially in linear regression. However, the focus here is on 'how to predict' more than on the general principles undergirding good prediction. We are aware that there are many statistical criteria which can be used to evaluate the performance of a given predictor; we briefly mention some of these in the contexts of Sections 2.2.4 and 5.1.2. However, we do not provide any further discussion on how to evaluate predictive techniques.

Even with our narrow focus on the construction of predictors there are numerous topics omitted here. The most serious omissions include predictors based on model averages and predictors based on data mining techniques such as trees, neural nets, and kernel-based methods. Likewise, we do not report on how predictive techniques interact with clustering, multitype data, or dimension reduction.

In the next section, Sec. 2, we review several familiar examples, some without explanatory variables and some with. Then, in Section 3, we look at prediction in a time series context. Generalizing this structure of data, we look at prediction in a longitudinal context in Section 4. In Section 5, we turn to survival analysis and give the predictive forms implied by the frequently used survival and hazard function approach. Finally, in Section 6 we discuss some of the implications of the predictive approach.

2. Familiar examples

Here we give examples of predictors in eight different contexts, namely, (1) IID data in the Frequentist case focusing on the normal and empirical distribution function; (2) IID data in the Bayes case, focusing on the normal; (3) Frequentist linear regression; (4) Bayes linear regression; (5) logistic regression; (6) quantile regression; (7) Bayes classification; and (8) Linear discriminant classification. The first two examples do not involve explanatory variables; the last six do.

2.1. No explanatory variables

2.1.1. Frequentist parametric case

The simplest prediction examples assume that X is not present, i.e., there are no covariates. So, let us consider the predictive analog to the estimation of a population mean when the variance is known, namely, using IID observations $y_1, \ldots, y_n$ to predict Y_{n+1}. The natural predictor is $\bar{y} = (1/n)\sum y_i$. Given that no model has been assumed, one might assume first and second moments of the Y_i's, setting $\mu = EY_i$ and $\sigma^2 = \mathrm{Var}(Y_i)$, and use standard inequalities (Chebyshev and triangle) to obtain

$$\begin{aligned} & P\left(|\bar{Y} - Y_{n+1}| \geq \sqrt{\frac{\sigma^2(1+(1/n))}{\tau}}\right) \\ & \leq \frac{\tau}{\sigma^2(1+(1/n))}\left((E|\bar{Y}-\mu|^2) + (E|\mu - Y_{n+1}|^2)\right) \\ & \leq \tau, \end{aligned} \tag{2.1}$$

for given $\tau > 0$. It is seen that the bound is only nontrivial when $\tau < 1$. For such a choice of τ, the natural Frequentist prediction interval (PI) is

$$\bar{Y} \pm \sigma\sqrt{\frac{(1+(1/n))}{\tau}}, \tag{2.2}$$

provided σ is known.

In the normal case $Y_i \sim N(\mu, \sigma^2)$ and direct calculation gives $\bar{Y} - Y_{n+1} \sim N(0, \sigma^2(1+\frac{1}{n}))$. The PI becomes $\bar{Y} \pm z_{1-\alpha/2}\sigma(1+(1/n))^{1/2}$ where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of $N(0,1)$.

Comparing the normal case with (2.2), we see that if we set $\tau = 1-\alpha$ to make the confidence levels the same, then the ratio of widths of the two PI's is

$$\frac{\sigma z_{1-\alpha/2}(1+(1/n))}{\sigma(1+(1/n))/(1-\alpha)} = z_{1-\alpha/2}(1-\alpha) \approx 2,$$

where $\approx$ means we have approximated $z_{1-\alpha/2} = 2$ and $\alpha = 0$ since these are close to commonly chosen values. This means that PI's from the general approach are around twice the width of the normal case. It is seen that Chebyshev's inequality in (2.1) is strong enough to give the $\mathcal{O}(1/n)$ rate seen in the normal, but too weak to give a ratio of widths shrinking to one with n.

The above bounds presume that $\bar{y}$ has been chosen as the appropriate predictor and this will usually be the case with normal data. However, if robustness were important, then a better choice than $\bar{y}$ for predicting Y_{n+1} would be $\mathsf{med}_i y_i$, the median of the y_i's, or possibly an estimate of the mode of Y_{n+1}. We ignore this here apart from noting that, asymptotically, the mean and the median are both $\mathcal{O}(1/\sqrt{n})$ even though the mean is more efficient. The argument to obtain (2.2) can be adapted to the median, and indeed, most other point estimators, possibly using the Hölder inequality rather than Cauchy-Schwarz.

In the case that σ is unknown, we have

$$\begin{aligned} P\left(\frac{|\bar{Y} - Y_{n+1}|}{\hat{\sigma}} \geq \frac{(1+(1/\sqrt{n}))}{\tau}\right) &\leq \frac{\tau}{(1+(1/\sqrt{n}))} E\left(\frac{1}{\hat{\sigma}}\right)|\bar{Y} - Y_{n+1}| \\ &\leq \frac{\tau}{(1+(1/\sqrt{n}))} \sqrt[u]{E\left(\frac{1}{\hat{\sigma}}\right)^u} \sqrt[v]{E|\bar{Y} - Y_{n+1}|^v}, \end{aligned} \tag{2.3}$$

where $\hat{\sigma}$ in an estimator of σ and $1/u + 1/v = 1$. When the Y_i's are normal, the usual estimator for σ is based on $s^2 = (1/(n-1))\sum(Y_i - \bar{Y})^2$ and it is well known that $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$. So, $\sigma^2/(s^2(n-1))$ is an Inverse-Chi-squared random variable with $n-1$ degrees of freedom which has mean $1/(n-3)$. Thus, $E(1/s^2) = (n-1)/[(n-3)\sigma^2]$. When $u = v = 2$, the triangle inequality gives that (2.3) becomes

$$\frac{\tau}{(1+(1/\sqrt{n}))} \frac{\sqrt{n-1}}{\sqrt{n-3}\sigma} \left(\sigma\left(1+\frac{1}{n}\right)^{1/2}\right) = \tau \frac{\sqrt{n-1}}{\sqrt{n-3}} \frac{\sqrt{1+(1/n)}}{(1+(1/\sqrt{n}))}. \tag{2.4}$$

Thus, the bound in (2.4) increases slightly over (2.1) due to the extra variability from $\hat{\sigma}$. So the PI's from (2.3) become

$$\bar{Y} \pm \hat{\sigma} \frac{\sqrt{n-1}\sqrt{1+(1/n)}}{\tau\sqrt{n-3}}, \tag{2.5}$$

where the factor $(1+1/\sqrt{n})^{-1}$ has been absorbed into τ, the prediction analog of confidence. An exact derivation using normality throughout rather than

Chebyshev's inequality in (2.3) can be found in [49] Chapter 2. When the underlying model is not normal, the width of the prediction interval is controlled by the last two factors in (2.3), and if the distribution of $1/\hat{\sigma}$ is too spread out, or too close to zero (or $|\bar{Y} - Y_{n+1}|$ has moments that are very high), then the τ would have to decrease to get the same confidence. Again, the PI is independent of the underlying model, apart from the first two moments.

If the data were paired, i.e., we have independent pairs (U_i, V_i) for $i = 1, \ldots, n$ and the task is to predict $U_{n+1} - V_{n+1}$, then the arguments can be applied to $Y_i = U_i - V_i$, yielding results analogous to (2.2) and (2.5).

It is worth commenting that if we set $\sigma = 1$ in the normal case, but do not take upper bounds as in (2.1) then there are two natural ways to get prediction intervals and they have different properties. The first is to recognize that $(\bar{Y} - Y_{n+1})/(1 + (1/n)) \sim N(0, 1)$ and therefore a $1 - \alpha$ PI is $\bar{Y} \pm z_{1-\alpha/2}\sqrt{(n+1)/n}$. The second is to use the estimate $\hat{\mu} = \bar{y}$ to give a distribution, namely $N(\bar{y}, 1)$, that can be used to predict Y_{n+1}. This gives

$$\hat{P}(\bar{Y} - z_{1-\alpha/2} \leq Y_{n+1} \leq \bar{Y} + z_{1-\alpha/2}) = 1 - \alpha,$$

where $\hat{P}(\cdot)$ is the probability assigning mass $\hat{P}(A) = N(\bar{y}, 1)(A)$ for a given set A. This means that the $1 - \alpha$ PI is $\bar{y} \pm z_{1-\alpha/2}$, slightly narrower than before – the factor on $z_{1-\alpha/2}$ is 1 rather than $(n+1)/n$. The difference is that the larger interval $\bar{Y} \pm z_{1-\alpha/2}\sqrt{(n+1)/n}$ includes the variability in the estimate of μ while the interval $\bar{Y} \pm z_{1-\alpha/2}$ is conditional on the use of the data via $\bar{y}$ to identify the prediction distribution. This distinction extends to the case that σ is unknown; see [49] example 2.2 p. 9. The normal example also extends to the q steps forward prediction; see [49], p. 10-11. Other examples can be derived and have similar properties, provided a pivotal quantity with a mathematically tractable density exists so that closed form expressions for the PIs can be derived.

Given a normal distribution with known mean, say zero, but unknown variance σ^2, it is well known that the sample variance s^2 of the observations $Y_1, \ldots, Y_n$ satisfies

$$\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1},$$

while the future observation Y_{n+1} has distribution $N(0, \sigma^2)$. Taking the ratio of the future observation Y_{n+1} and the sample standard deviation cancels the σ and gives a Student's t-distribution with $n-1$ degrees of freedom, i.e., $Y_{n+1}/s \sim t_{n-1}$. Solving for Y_{n+1} gives the prediction distribution st_{n-1}, from which PIs can be found. Notice that this prediction distribution gives slightly larger PIs than just using a $N(0, s^2)$ because the t-distributions have heavier tails than the normal. This is necessary for the interpretation of the confidence level $1 - \alpha$ in finite samples but the two are asymptotically equivalent.

2.1.2. Frequentist non-parametric case

The assumption of normality can be relaxed so that a nonparametric approach to obtaining PIs can be given based on the empirical distribution function (EDF)

$\hat{F}_n(\cdot)$. There are two versions of this. The first is to take an interval based on $\hat{F}^{-1}$ and expand it using the uncertainty in $\hat{F}$ as bounded by the Smirnov theorem. The second is to take endpoints for an interval using percentiles of $\hat{F}$ and enlarge it by invoking a central limit theorem for those percentiles. We will call on these results in Section 5 where we obtain estimates $\hat{S}_n$ of the survival function S.

To begin, note that a confidence level for a PI can be found by using a version of the Kolmogorov-Smirnov theorem. Theorem 4 in [43] establishes

$$\lim_{n\to\infty} P(\sqrt{n}\sup_y |\hat{F}_n(y) - F(y)| < \epsilon) = (1 - e^{-2\epsilon^2}).$$

Now, letting $\epsilon > 0$ and setting $A_{\epsilon,n} = \{\sup_y |\hat{F}_n(y) - F(y)| < \epsilon/\sqrt{n}\}$ to be the 'good' set, we have that for given $\alpha > 0$,

$$\begin{aligned}
P(Y_{n+1} \in\ & [\hat{F}_n^{-1}(\alpha/2), \hat{F}_n^{-1}(1-\alpha/2)]) \\
=\ & P(F(Y_{n+1}) + (\hat{F}_n(Y_{n+1}) - F(Y_{n+1})) \in [\alpha/2, 1-\alpha/2]) \\
\leq\ & P(F(Y_{n+1}) \in [\alpha/2 - |\hat{F}_n(Y_{n+1}) - F(Y_{n+1})|, \\
& \qquad 1-\alpha/2 + |\hat{F}_n(Y_{n+1}) - F(Y_{n+1})|]) \\
\leq\ & P(F(Y_{n+1}) \in [\alpha/2 - |\hat{F}_n(Y_{n+1}) - F(Y_{n+1})|, \\
& \qquad 1-\alpha/2 + |\hat{F}_n(Y_{n+1}) - F(Y_{n+1})|] \cap A_{\epsilon,n}) + P(A^c_{\epsilon,n}) \\
\leq\ & P(F(Y_{n+1}) \in [\alpha/2 - \epsilon/\sqrt{n}, 1-\alpha/2 + \epsilon/\sqrt{n}]) + \eta + e^{-2\epsilon^2} \qquad (2.6) \\
=\ & 1 - \alpha + \frac{2\epsilon}{\sqrt{n}} + \eta + e^{-2\epsilon^2}, \qquad (2.7)
\end{aligned}$$

for $n \geq N$. To choose N, note that Feller's Theorem ensures there is an N large enough that $|P(A_{\epsilon,n}) - (1 - e^{-2\epsilon^2})| \leq \eta$ for pre-assigned $\eta > 0$. Hence, using $P(A^c_{\epsilon,n}) = 1 - P(A_{\epsilon,n})$ we get the bound $|P(A^c_{\epsilon,n}) - e^{-2\epsilon^2}| < \eta$ for $n > N$ as used in (2.6). It is seen that as $n \to \infty$ we can let ϵ and η go to zero so that the asymptotic confidence of the PI is $1 - \alpha$. (To get the equality at (2.7) we also used the fact that $F(Y_{n+1}) \sim \text{Unif}[0, 1]$.)

A second way to look at prediction from the EDF is to observe that the rate at which the confidence of a PI of the form $[\hat{F}_n^{-1}(\alpha/2), \hat{F}_n^{-1}(1-\alpha/2)]$ approaches $1 - \alpha$ is the usual $\sqrt{n}$ rate. To see this, recall the asymptotic normality of quantiles. In fact, [82] Chap. 4, Sec. 1 shows that for any $\alpha \in (0, 1)$,

$$\hat{F}_n^{-1}(\alpha) \to N\left(F^{-1}(\alpha), \frac{\alpha(1-\alpha)}{nF'(F^{-1}(\alpha))^2}\right), \qquad (2.8)$$

weakly as $n \to \infty$. Now, using (2.8), asymptotic $100\gamma/2\%$ lower and $100(1 - \gamma/2)\%$ upper confidence bounds are given as

$$F^{-1}(\alpha/2) \geq F_n^{-1}(\alpha/2) - z_{\gamma/2}\frac{\sqrt{\alpha/2(1-\alpha/2)}}{\sqrt{n}F'(F^{-1}(\alpha/2))}$$

and

$$F^{-1}(1-\alpha/2) \leq F_n^{-1}(1-\alpha/2) - z_{1-\gamma/2}\frac{\sqrt{(1-\alpha/2)(\alpha/2)}}{\sqrt{n}F'(F^{-1}(1-\alpha/2))},$$

where $z_{\gamma/2}$ is the $100\gamma/2$ quantile of a $N(0,1)$. That is, we obtain a $1-\alpha$ confidence PI for Y_{n+1} as

$$\left[F_n^{-1}(\alpha/2) - z_{\gamma/2}\frac{\sqrt{\alpha/2(1-\alpha/2)}}{\sqrt{n}F'(F^{-1}(\alpha/2))}\right., \\ \left. F_n^{-1}(1-\alpha/2) - z_{1-\gamma/2}\frac{\sqrt{(1-\alpha/2)(\alpha/2)}}{\sqrt{n}F'(F^{-1}(1-\alpha/2))}\right]. \tag{2.9}$$

The intervals in (2.7) and (2.9) are roughly comparable; the difference is that the $\sqrt{n}$ rate is identified in (2.9). Note that the argument leading to (2.9) treats the upper and lower bounds for $\alpha/2$ and $1-\alpha/2$ separately. In fact, they are dependent and a joint asymptotic normality result for $(\hat{F}_n^{-1}(\alpha/2), \hat{F}_n^{-1}(1-\alpha/2))$ can be given; see [82]. The covariances are of the same order as in (2.9) but the constants change slightly. In addition, PI bounds can also be found using the Kolmogorov-Smirnov bound for a distribution function; see [31], p. 395.

2.1.3. Bayesian parametric case

The simplest Bayesian version of this is to assume that the Y_i's are drawn IID from some density $p(\cdot|\theta)$ where Θ is a random variable with outcomes θ and density $w(\theta)$. Then, the Bayesian forms a posterior density $w(\theta|y^n)$ for θ given the data, i.e.,

$$w(\theta|y^n) = w(\theta)p(y^n|\theta)/m(y^n)$$

where

$$m(y^n) = \int w(\theta)p(y^n|\theta)d\theta. \tag{2.10}$$

Then $w(\theta|y^n)$ describes the post-data variability of the parameter. So, the Bayesian writes the predictive density

$$m(Y_{n+1} = y_{n+1}|y^n) = \frac{m(y^{n+1})}{m(y^n)} = \int p(y_{n+1}|\theta)w(\theta|y^n)d\theta \tag{2.11}$$

and uses it to form a highest posterior density region $R(\alpha) = R_{n+1}(y^n;\alpha)$ with probability $1-\alpha$ in $m(\cdot|y^n)$ that satisfies

$$\int_{R(\alpha)} m(y_{n+1}|y^n) = 1-\alpha. \tag{2.12}$$

Implementing this may be difficult in closed form but it can be done computationally quite readily. Note that (2.12) depends on the prior w and uses a conditional probability given the data.

Bayesian point and interval predictors can also be found – but the PI's often are Frequentist in that they are defined using P_{θ_T}. One choice for a Bayesian

point predictor is

$$\hat{Y}_{n+1} = \int y_{n+1} m(y_{n+1}|y^n) dy_{n+1} = \int E_\theta(Y_{n+1}) w(\theta|y^n) d\theta \to \mu = E_{\theta_T}(Y),$$

as $n \to \infty$ when θ_T is taken as the true value of θ. This leads to Frequentist derived PIs by choosing $\tau > 0$ and, analogous to (2.1), writing

$$\begin{aligned} P_{\theta_T} \ (|\hat{Y}_{n+1} - Y_{n+1}| > \tau) \\ &\leq \frac{1}{\tau^2} E_{\theta_T} \left(\int |Y_{n+1} - E_\theta(Y)| w(\theta|y^n) d\theta \right)^2 \\ &\leq \frac{1}{\tau^2} E_{\theta_T} \int |Y_{n+1} - E_\theta(Y)|^2 w(\theta|y^n) d\theta \\ &\leq \frac{1}{\tau^2} \sqrt{E_{\theta_T}(Y_{n+1} - \mu)^2} + \frac{1}{\tau^2} E_{\theta_T} \int |\mu - E_\theta(Y)|^2 w(\theta|y^n) d\theta \end{aligned} \quad (2.13)$$

using Markov, the triangle inequality and Cauchy-Schwartz. The first term in (2.13) gives σ^2/τ^2. To evaluate the second term, note that in the one-dimensional case a Taylor expansion gives

$$E_\theta(Y_{n+1}) - E_{\theta_T}(Y) = \frac{1}{2}(\theta - \theta_T) \frac{\partial E_\theta(Y)}{\partial \theta}\bigg|_{\theta=\theta_T} + o(\|\theta - \theta_T\|), \quad (2.14)$$

where the little-o bound holds on a small neighborhood of θ_T where $w(\theta|y^n)$ concentrates. Also, recall that posterior normality gives

$$w(\theta|y^n) \approx \frac{\sqrt{n}}{\sqrt{2\pi I(\theta_T)}} e^{-nI(\theta_T)^{-1}(\theta-\theta_T)^2/2}.$$

So, the second term in (2.13) is approximately

$$\begin{aligned} E_{\theta_T} \int \frac{1}{4}(\theta - \theta_T)^2 \ \left| \tfrac{\partial E_\theta(Y)}{\partial \theta} \right|^2_{\theta=\theta_T} \ \frac{\sqrt{n}}{\sqrt{2\pi I(\theta_T)}} e^{-nI(\theta_T)^{-1}(\theta-\theta_T)^2/2} d\theta \\ = \frac{1}{4} \left| \frac{\partial E_\theta(Y)}{\partial \theta} \right|^2_{\theta=\theta_T} \frac{I(\theta_T)}{n}, \end{aligned} \quad (2.15)$$

apart from the factor $1/\tau^2$, by a standard Laplace's approximation argument. Now, (2.13) becomes

$$P_{\theta_T}(|\hat{Y}_{n+1} - Y_{n+1}| > \tau) \leq \frac{1}{\tau} \left(\sigma^2 + B(\epsilon, \theta_T) \frac{I(\theta_T)}{n} \right)^{1/2} \quad (2.16)$$

where the factor $B(\epsilon, \theta_T)$ includes the absolute first derivative from (2.15) and a bound $(1 + \epsilon)$ from a Laplace's approximation (where ϵ is the width of the neighborhood used in the Laplace approximation). Note that this ϵ can go to zero as $n \to \infty$. The technique giving the bound in (2.16) generalizes to give

an analogous bound when θ is a general finite dimensional parameter. It is seen that (2.16) is parallel to (2.1) and gives PIs of the form (2.2).

It is seen from (2.12) that the key to Bayes prediction is obtaining a closed form for the mixture $m(\cdot)$ of densities in (2.10). This is (relatively) easy when the prior is conjugate to the likelihood since the posterior can be obtained in closed form. The case of a normal variable with a normal prior is particularly easy. If $w(\theta)$ is $N(\mu, \tau^2)$ and $Y_i|\theta \sim N(\theta, \sigma^2)$ where μ, τ, and σ are known, then $\bar{Y} \mid \theta \sim N(\theta, \sigma^2/n)$ and

$$m(\bar{y}) \quad = \quad \frac{\sqrt{n}}{\sigma\tau\sqrt{2\pi\rho}} e^{-\frac{(\mu-\bar{y})^2}{2(\sigma^2/n+\tau^2)}}, \tag{2.17}$$

the density of a $N(\mu, \sigma^2/n+\tau^2)$, where $\rho = 1/(\sigma^2/n) + 1/\tau^2$. So, $w(\theta|y^n)$ is the density of a

$$N(\frac{\sigma^2/n}{\sigma^2/n+\tau^2}\mu + \frac{\tau^2}{\sigma^2/n+\tau^2}\bar{y}, \frac{1}{\rho}) = N(E(\Theta|y^n), \mathrm{Var}(\Theta|y^n)).$$

The posterior mean is $E(\Theta|y^n) = \tau^2/(\sigma^2/n+\tau^2)\bar{y} + (\sigma^2/n)/(\sigma^2/n+\tau^2)\mu$ and the posterior variance is $1/\rho = \tau^2\sigma^2/(n\tau^2+\sigma^2) = \mathcal{O}(1/n)$. Now, it can be directly verified that $m(y_{n+1}|y^n)$ is the density of a

$$N(E(Y_{n+1}|Y^n = y^n), \mathrm{Var}(Y_{n+1}|Y^n = y^n))$$

and $E(Y_{n+1}|Y^n = y^n) = E(\Theta|Y^n = y^n)$, so that

$$\mathrm{Var}(Y_{n+1}|Y^n = y^n) = \sigma^2 + \mathrm{Var}(\Theta|Y^n = y^n).$$

So, one choice for $R(\alpha)$ is $E(\Theta|Y^n = y^n) \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(Y_{n+1}|Y^n = y^n)}$ for fixed y^n. Note that intervals of this form asymptotic have confidence $(1-\alpha)$ in the predictive distribution $M(\cdot|y^n)$ for Y_{n+1}, i.e.,

$$M\left(E(\Theta|Y^n = y^n) - z_{1-\alpha/2}\sqrt{\mathrm{Var}(Y_{n+1}|Y^n = y^n)} \le Y_{n+1}\right.$$
$$\left. \le E(\Theta|Y^n = y^n) + z_{1-\alpha/2}\sqrt{\mathrm{Var}(Y_{n+1}|Y^n = y^n)|y^n)}\right) = 1-\alpha.$$

So, if the data set y^n is held fixed, the variability in the data only affects the prediction via the model. If one did not use $M(\cdot|y^n)$ to get prediction intervals, the properties of $E(\Theta|Y_n = y_n)$ as a predictor for Y_{n+1} would change. For instance, one could obtain PI's for Y_{n+1} based on $E(\Theta|Y_n = y_n)$ using the unconditional $M(\cdot)$ or P_{θ_T} for Y^{n+1}. In these cases, one would expect the $1-\alpha$ PI's to be larger due to replacing the fixed y^n with the random Y^n. On the other hand, a prior often corresponds to having extra data shrinking the PI's. So, there will be a tradeoff between the extra information in the prior and the extra variability in Y^n if one finds PI's using an unconditional distribution.

Even if σ^2 is not known, the calculations can be done to obtain an explicit form for $m(y_{n+1}|y^n)$. Suppose $Y_i \sim N(\mu, \sigma^2)$ and $w(\mu, \sigma^2) \propto 1/\sigma^2$. Then,

$$\begin{aligned} w(\mu, \sigma^2|\bar{y}) \quad &\propto \quad w(\mu, \sigma^2)p(y^n|\mu, \sigma^2) \\ &\propto \quad \frac{1}{\sigma^2}\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-(1/2\sigma^2)[(n-1)s^2+n(\bar{y}-\mu)^2]}, \end{aligned}$$

where $s^2 = (1/(n-1))\sum_i (y_i - \bar{y})^2$. The posterior distribution of σ^2 given $\bar{y}$ is

$w(\sigma^2|\bar{y}) \sim$ Scaled Inverse -$\chi^2(n-1, s^2)$ = Inverse-Gamma$((n-1)/2, (n-1)s^2/2)$

and $w(\theta|\sigma^2, \bar{y})$ is the density of a $N(E(\Theta|y^n), \mathrm{Var}(\Theta|y^n))$. Writing $\mu = E(\Theta|y^n)$ and $\tau^2 = \mathrm{Var}(\Theta|y^n)$ in (2.11) we have

$$\begin{aligned} m(y_{n+1}|y^n) &= \int p(y_{n+1}|\theta,\sigma) w(\theta|\sigma^2,\bar{y}) w(\sigma^2|\bar{y}) d\theta d\sigma^2 \\ &= \int \phi_{\mu,\sigma^2+\tau^2}(y_{n+1}) w(\sigma^2|\bar{y}) d\sigma^2, \end{aligned} \tag{2.18}$$

where $\phi_{a,b}$ indicates the $N(a,b)$ density, and it can be verified that $m(y_{n+1}|y^n)$ has the density of a $t_{n-1,\bar{y},s^2(1+1/n)}$ random variable, which is a special case of (2.20) below. Here $t_{n-1,\bar{y},s^2(1+1/n)}$ is the Student t distribution with degrees of freedom $n-1$, location $\bar{y}$, and scale $s^2(1+1/n)$.

Taken together the above derivations for PI's in the normal case show that the standard results from normal estimation theory carry over to analogs for prediction. This means that one can regard prediction as the central goal instead of parameter estimation – without loss. Even better, taking upper bounds makes the PI's somewhat independent of the specific model without much increase in the width of the PI's. However, it must be remembered that these are purely methods of construction of PI's. In the presence of nontrivial model mis-specification the PI's would have to be validated on new data.

2.2. Explanatory variables present

In this subsection, we merely quote the results since techniques for model selection, parameter estimation, and model validation are well known and can be found in most references. Thus, we list and explain the forms of predictors with minimal description.

2.2.1. Fixed effects linear regression

Another familiar example of prediction comes from linear regression. Recall the model:

$$Y_i = X_i\beta + \epsilon_i$$

where $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ is the parameter vector for explanatory variables $X_i = (x_{1i}, \ldots, x_{di}) \in \mathbb{R}^d$, and the data (y_i, x_i) for $i = 1, \ldots, n$ is available where the noise terms ϵ_i are taken to be IID $N(0, \sigma^2)$ outcomes. If the goal is to predict Y_{n+1} for a new design point X_{n+1} then the point prediction is given by $\hat{Y}_{n+1}(X_{n+1}) = X_{n+1}\hat{\beta}$ where $\hat{\beta}$ is the least squares estimate for β. The prediction interval is given by

$$\hat{Y}_{n+1} \pm t_{1-\alpha/2;n-d}\hat{\sigma}\sqrt{1 + X_{n+1}^T(X^TX)^{-1}X_{n+1}},$$

where X denotes the $n \times d$ design matrix, $n > d$, and $\hat{\sigma}$ is the root of the residual sum of squares divided by its degrees of freedom. If the distribution of ϵ is not normal, the form of the interval will change, but the point prediction does not. In the special case that $d = 2$ where $X_{1,i} = 1$ and $X_{2,i} = x_i$, so that $\bar{x} = \bar{x}_2$, the form of the interval is

$$\hat{Y}_{n+1} \pm t_{1-\alpha/2;n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

If $x_{n+1} = \bar{x}$, as may be approximately true in designed experiments, then the variance reduces to $(1 + (1/n))$ as in the normal case without covariates.

2.2.2. Bayesian linear regression

The Bayesian analog to fixed effects regression has been worked out; see [17, 48]. In hierarchical form, the Bayesian model with noninformative priors is

$$\begin{aligned}(Y_1, \ldots, Y_n|\beta, \sigma^2, X) &\sim N(X\beta, \sigma^2 I_n)\\ w(\beta, \sigma^2) &\propto \frac{1}{\sigma^2} \quad \text{for} \quad \sigma > 0, \beta \in \mathbb{R}^d.\end{aligned} \tag{2.19}$$

(A related Frequentist analysis is discussed in Sec. 4.1.) The posterior for β given σ^2 is

$$(\beta|\sigma^2, y^n, X) \sim N((X^TX)^{-1}X^Ty^n, \sigma^2(X^TX)^{-1}),$$

and the marginal posterior for σ^2 is

$$(\sigma^2|y^n, X) \sim \text{Scaled Inverse-}\chi^2_{n-d,s^2}.$$

The scale factor s^2 is the residual squared error divided by its degrees of freedom, $s^2 = (y^n - X(X^TX)^{-1}X^Ty^n)^T(y^n - X(X^TX)^{-1}X^Ty^n)/(n-d)$. It can be verified that the marginal posterior distribution of $(\beta|y^n)$ is

$$(\beta|y^n, X) \sim t^d(n-d, (X^TX)^{-1}X^Ty^n, s^2),$$

where $t^d(n-d, (X^TX)^{-1}X^Ty^n, s^2)$ is the d-dimensional Student t distribution with degrees of freedom $n-d$, location $(X^TX)^{-1}X^Ty^n$, and scale s^2. To make a prediction for Y_{n+1} at a new value X_{n+1}, a Bayesian might use the predictive distribution. If σ is given, the predictive distribution is normal with mean

$$\hat{Y}(X_{n+1}) = X_{n+1}(X^TX)^{-1}X^Ty^n$$

and variance

$$\text{Var}(\hat{Y}(X_{n+1})|\sigma^2, y^n) = \sigma^2(1 + X_{n+1}(X^TX)^{-1}X^T_{n+1}),$$

paralleling the Frequentist case. If σ is unknown, the predictive distribution is

$$Y_{n+1}|y^n, X, X_{n+1} \sim t(n-d, X_{n+1}(X^TX)^{-1}X^Ty^n, s^2(1+X_{n+1}(X^TX)^{-1}X_{n+1}^T)), \quad (2.20)$$

when σ is integrated out.

Other choices for the priors on the parameters are possible as well. For instance, it is not very hard to derive closed form expressions for posterior densities when $p(\sigma^2)$ is an Inv-Gamma(a,b) density and $\beta \sim N(\mu, \sigma^2 V)$, i.e., when conjugate priors are assigned to the parameters, see [4] among others. Closed form expressions can be given for the posterior $p(\beta, \sigma^2|y^n, X)$ (normal-Inverse-Gamma due to conjugacy), the marginal posteriors $p(\sigma^2|y^n, X)$ (an Inverse-Gamma) and $p(\beta|y^n, X)$ (a multivariate t), and most importantly for $p(y_{n+1}|y^n, X, X_{n+1})$ (a multivariate t).

Rather than presenting the conjugate prior case explicitly, we consider the use of Zellner's g-prior. That is, we retain the likelihood portion of (2.19) but change from the noninformative prior to the informative Zellner's g-prior for β, see [99]. This is given by

$$(\beta \mid \sigma^2, X) \sim N(\beta_0, g\sigma^2(X^TX)^{-1}) \quad (2.21)$$

in which g is a hyperparameter. Zellner's g-prior is motivated by the form of the variance in the noninformative case and also by trying to control the collinearity among the covariates in X. Using (2.21) leads to the conditional posterior

$$(\beta|y^n, \sigma^2, X, g) \sim N(\frac{\beta_0 + g\hat{\beta}}{g+1}, \frac{g\sigma^2}{g+1}(X^TX)^{-1}) \quad (2.22)$$

in which $\hat{\beta} = (X^TX)^{-1}Xy^n$ is the MLE. The marginal likelihood under (2.21) when $\beta_0 = 0$ can be derived and is

$$f(y^n|\sigma^2, X, g) = \frac{1}{(2\pi\sigma^2)^{n/2}(1+g)^{-d/2}} e^{-(y^n)^Ty^n/2\sigma^2 + (g/(2\sigma^2(1+g)))\hat{\beta}^TX^TX\hat{\beta}}.$$

The marginal posterior is a location-scale t-distribution,

$$(\beta \mid y^n, X, g) \sim t_d(n-p, \frac{\beta_0 + g\hat{\beta}}{g+1}, \frac{g(s^2 + (\beta_0 - \hat{\beta})^TX^TX(\beta_0 - \hat{\beta}))}{n(g+1)^2}(X^TX)^{-1}),$$

and one form of the posterior predictive distribution is

$$(Y_{n+1} \mid \sigma^2, X_{n+1}, y^n, X, g) \sim N(\frac{X_{n+1}(\beta_0 + g\hat{\beta})}{g+1}, \sigma^2(1 + (g/(g+1))X_{n+1}^T(X^TX)^{-1}X_{n+1})).$$

A closed-form derivation to obtain $(Y_{n+1}|X_{n+1}, X)$ using Zellner's g-prior does not seem to have been worked out, in part because there is controversy about what hyperprior to assign to g or whether to choose a good value $\hat{g}$ in some other way and use $(Y_{n+1}|X_{n+1}, X, \hat{g})$. However, the usual priors on σ – Inverse-Gamma or the noninformative $1/\sigma^2$ – are frequently still used in this case.

2.2.3. Logistic and quantile regression

Two other basic predictors that are important to consider here are logistic regression and quantile regression.

In the simplest logistic regression model, the log of the odds ratio is expressed as an affine function of a single explanatory variable x:

$$\log \frac{P(Y=1|\beta_0,\beta_1,x)}{1-P(Y=1|\beta_0,\beta_1,x)} = \beta_0 + \beta_1 x; \tag{2.23}$$

extensions to include more explanatory variables are similar. The kind of data that is available is usually of the form (Y_i, x_i) for $i = 1,\ldots,n$. In the simplest case, Y and x are binary, i.e., assume values zero-one. In this case, β_0 and $\beta_0+\beta_1$ in (2.23) are the values of the true log odds (for $x=0$ and for $x=1$).

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained in a variety of standard ways (including Bayesian; see [24], Chap. 8) and the appropriateness of the model is usually assessed in a variety of standard ways. Here, the point is to estimate the probabilities of events by

$$\hat{P}(Y=1|\hat{\beta}_0,\hat{\beta}_1,X_{new}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_{new}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_{new}}}.$$

Clearly, if $\hat{P}(Y=1|\hat{\beta}_0,\hat{\beta}_1,x_{new})$ is greater than $1/2$, one is led to predict $Y=1$ for x_{new} and otherwise to predict $Y=0$; values of Y in $(0,1)$ may also be used but this case is not handled here. Note that strictly speaking, $\hat{P}(Y=1|\hat{\beta}_0,\hat{\beta}_1,x_{new})$ estimates $P(Y=1|\beta_0,\beta_1,x_{new})$ since it is not a random variable. However, the estimate of the probability leads naturally to a predictor for Y by choosing the value of Y which has higher probability. Note that x can be multidimensional and continuous; Y is typically discrete but needn't be binary.

When a regression function assumes discrete values say $0,1,\ldots,K-1$ the problem is usually called classification rather than regression and the discrete values are called class labels. In these cases, the regression function is called a classifier. The values of x for which $Y=k$ is called the k-th subpopulation or k-th class, $k=0,\ldots,K-1$. In this sort of problem the error term is discrete not continuous; this represents the possibility that one may predict the wrong class for a given x. More generally, one may obtain a classifier from an estimated regression function $\hat{Y}$ that is continuous by assigning x_{n+1} the class label that is closest to $\hat{Y}(x)$. Thus, logistic regression is one of the simplest ways to do binary ($K=2$) classification.

The last of the regression-based predictors that will be familiar to most people derives from quantile regression (QR); see [64] for an introductory exposition, and [65] for the original contribution. We comment that QR has been extended to neural nets [89] and to forests (i.e., tree-based ensembles) [70]. The simple version of the problem can be succinctly stated as follows. Let $\mu(x,\beta)$ be a smooth class of functions where x and β are d dimensional. Assume also that

data of the form $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ are available. Then find

$$\hat{\beta}_\tau = \arg\min_\beta \sum_{i=1}^{n} \rho_\tau(y_i - \mu(x_i, \beta)), \tag{2.24}$$

where ρ_τ is the modification of $|\cdot|$ that has slope K_L on the left of zero and slope K_R on the right of zero and $\tau = K_L/(K_L + K_R)$. (This means that the Bayes action under loss ρ_τ is the τ-th quantile.) This procedure is thought to be good when the conditional distribution of $(Y|X)$ has thick tails, is asymmetric, or is not unimodal. If desired, $y_i - \mu(x_i, \beta)$ can be replaced by $(y_i - \mu(x_i, \beta))/\sigma$ and a factor $1/\sigma$ put in front of the summation.

The simplest choice of μ in (2.24) is linear, i.e., $\mu(x, \beta) = \beta^T x$. Now, (2.24) can be written

$$\hat{\beta}_\tau = \arg\min_\beta \left[\sum_{i \in \{i: y_i \geq \beta x_i\}} \tau |y_i - \beta^T x_i| + \sum_{i \in \{i: y_i < \beta x_i\}} (1-\tau)|y_i - \beta^T x_i| \right].$$

There are several ways to solve this optimization problem computationally; this has been implemented in R, Matlab, and SAS, for example, but is not discussed here. The point is that if $\hat{\beta}_\tau$ is found for say $\tau = .05$ and .95, then we have a PI for Y_{n+1} from $\hat{\beta}_{.05} x_{n+1}$ and $\hat{\beta}_{.95} x_{n+1}$, if the variability in (and dependence between) $\hat{\beta}_{.05}$ and $\hat{\beta}_{.95}$ is ignored. To include the variability in $\hat{\beta}_{.05}$ and $\hat{\beta}_{.95}$, we would have to look at, for instance, a .975 lower confidence bound for $\min(\hat{\beta}_{.05} x_{n+1}, \hat{\beta}_{.95} x_{n+1})$ and an upper .975 confidence bound on $\max(\hat{\beta}_{.05} x_{n+1}, \hat{\beta}_{.95} x_{n+1})$. Note this procedure is robust since it is based on percentiles i.e., rankings, rather than sums. For more details on the use and properties of QR, see [63].

Bayesian QR has also been developed; it originates in [93], but has seen rapid development since. Following [93], the likelihood function for $f_\tau(y^n \mid \beta)$ is

$$L_\tau(\beta \mid y^n) = \tau^n (1-\tau)^n e^{-\sum_{i=1}^n \rho_\tau(y_i - x_i^T \beta)}, \tag{2.25}$$

where τ and ρ_τ are as in (2.24). Essentially, (2.25) means that for each τ, each data pair (y_i, x_i) follows an asymmetric Laplace density $f_\tau(y_i|\beta) = \tau(1-\tau)e^{-\rho_\tau(y_i - x_i^T\beta)}$ (where $\sigma = 1$). Often, the components of β are assigned independent improper uniform prior distributions (a standard conjugate prior distribution is not available for the quantile regression formulation).

As shown in [93], the resulting posterior distribution of β is proper, and MCMC methods may be used to give numerical approximations for the posteriors of unknown parameters. If we denote the prior on β by $w(\beta)$ then given $(y_1, \ldots, y_n)$ we must approximate the posterior distribution of β

$$w_\tau(\beta \mid y_1, \ldots, y_n) \propto L_\tau(\beta \mid y_1, \ldots, y_n) w(\beta).$$

Given such an approximation, the predictive posterior distribution is given by

$$m_\tau(y_{n+1} \mid y_1, \ldots, y_n) = \int L_\tau(y \mid \beta) w_\tau(\beta \mid y_1, \ldots, y_n) d\beta, \tag{2.26}$$

from which PIs can be determined. That is, closed form expressions for PIs are not available unless closed form expressions for $w_\tau(\beta|y^n)$ are available and typically they're not. So, we can only form predictors computationally.

2.2.4. The Bayes classifier

The Bayes classifier identifies the most likely class for a given observation by modeling the distribution of the explanatory variables. That is, instead of regarding $x = (x_1, \ldots, x_d)$ as deterministic, we regard it as $X = (X_1, X_2, \ldots, X_d)$ and modeling it across classes. In Bayes classification, the response Y represents the class of an observation, i.e., $Y = 1, \ldots, K$ where K is the number of classes. Bayes theorem gives us the probability of an observation falling into the kth class, i.e.,

$$P(Y = k \mid X_1, X_2, \ldots, X_d) = \frac{P(Y = k)P(X_1, X_2, \ldots, X_d \mid Y = k)}{P(X_1, X_2, \ldots, X_d)}, \quad (2.27)$$

where the denominator can be expressed as

$$P(X_1, X_2, \ldots, X_d) = \sum_{k=1}^{K} P(Y = k)P(X_1, X_2, \ldots, X_d \mid Y = k). \quad (2.28)$$

We focus on the numerator as the denominator does not depend on k. The probability $P(Y = k)$ for $k = 1, \ldots, K$ represents the prior on Y, or what we believe a priori to be the proportions of each class in the population. Given an outcome of the explanatory vector $x = (x_1, \ldots, x_d)$ the Bayes classifier is the mode of (2.27),

$$\arg\max_y P(y)P(x_1, \ldots, x_d \mid Y = y). \quad (2.29)$$

In other words, given x, our point prediction, $\hat{y}$, of the corresponding y is given by (2.29). The Bayes classifier minimizes the expected cost of misclassification, i.e., the Bayes risk, under 0-1 loss and more general loss functions; see [26] Chapter 5.2. In practice, given n data points, one estimates the optimal Bayes rule by estimating the probabilities or densities in (2.27). For example, if $K = 2$, $P(Y = k \mid X)$ can be obtained by estimating the densities for $(X \mid Y = 1)$ and $(X \mid Y = 2)$ and $P(Y = k)$ can be obtained by using the sample proportions. These estimates can then be used in (2.28) to obtain the denominator of (2.27).

Note that both logistic regression and the Bayes classifier provide probabilities of membership in a given, predefined class. These probabilities can be used as the basis of our predictions of class membership. Such predictions, for $K = 2$, are often evaluated in terms of sensitivity Se and specificity Sp [3], defined as

$$\begin{aligned} Se &= P(\hat{y} = 1|y = 1) \\ Sp &= P(\hat{y} = 0|y = 0). \end{aligned}$$

In medical contexts sensitivity is interpreted as the true positive rate (TPR), while specificity is interpreted as the true negative rate (TNR). Predictors are

often evaluated in terms of their sensitivity and specificity, although the relative importance assigned to each of these components varies with the purpose of the predictor and the costs of misclassification. For further discussion of these and other measures see [3] and [88].

2.2.5. Linear discriminant analysis

The idea behind discriminant analysis for classification is that given a response Y that assume values in $0, 1, \ldots, K-1$, one derives K functions $\delta_k(\cdot)$, $k = 0, 1, \ldots, K$ such that $\delta_k(x)$ can be used to assess how representative each class is for a given x. Given data (Y_i, x_i), $i = 1, \ldots, n$ one estimates the δ_k's by $\hat{\delta}_k$'s. The discriminant classifier predicts class $\hat{Y}(x_{n+1}) = k^*$ for x_{n+1} where

$$k^* = \arg\max_k \hat{\delta}_k(x_{n+1}). \tag{2.30}$$

Here we will focus on the linear discriminant classifiers pioneered by Fisher in the 1930's, see [44], and now referred to collectively as Fisher's linear discriminant analysis (LDA). In LDA we assume, as in the last section, that x is a random variable X and that $(X \mid Y = k) \sim N(\mu_k, \Sigma)$ for $k = 0, \ldots, K-1$. Each class has the same covariance matrix Σ assumed to be full rank. Then the optimal predictor for $(Y \mid X = x)$ is

$$\begin{aligned}
\hat{Y} &= \arg\max_k P(Y = k \mid X = x) \\
&= \arg\max_k \left[-\log((2\pi)^{p/2} \mid \Sigma \mid^{1/2}) - \frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) \right] \\
&= \arg\max_k \left[-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) \right] \\
&= \arg\max_k \left[x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k - \frac{1}{2}x^T\Sigma^{-1}x \right] \\
&= \arg\max_k \left[x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k \right].
\end{aligned} \tag{2.31}$$

The expression inside the brackets in (2.31) is Fisher's linear discriminant function (LDF), a particular choice for $\delta_k(\cdot)$. The boundary between any two classes j and l is $\{x : \delta_k(x) = \delta_l(x)\}$. Indeed, it can be seen from (2.31) that Fisher's LDF specifies a plane in d dimensions which partitions the space of explanatory variables since the expression inside the brackets depends linearly on x.

Analogous to Bayes classification, LDA is implemented in practice by estimating the parameters in (2.31). Usually, the standard estimates $\hat{\Sigma}$ and $\hat{\mu}_k$ for $k = 0, \ldots, K-1$ are used, i.e.,

$$\hat{\mu}_k = \sum_{i:y_i(x_i)=k} x_i/n_k \text{ and } \hat{\Sigma} = \sum_{k=1}^{K} \sum_{i:y_i(x_i)=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(n-K), \tag{2.32}$$

where n_k is the number of observations in class k and n is the total number of observations.

3. Time series

In the time series literature prediction is usually called forecasting because the goal is to make statements about events that have not yet been observed in time. That is, i is a time. (Prediction is more general: One may predict, say, the temperature at a location based on a nearby location at the same time so that i may be a point on a grid where we have a measurement.) Regardless of the terminology, time series is a vast subject and no survey can do it justice. Here, we focus on a narrow version of the Box-Jenkins (BJ) methodology for prediction in weakly stationary processes, usually mean zero. Weakly stationary means that the first two moments are the same for all n and autocovariances of all orders are the same, i.e., for any $k \geq 0$, the $\gamma(k) = \mathrm{Cov}(Y_n, Y_{n+k})$s are independent of n. The BJ method for prediction in time series has four steps: 1) Model Class Identification; 2) Parameter Estimation; 3) Internal Validation; and 4) Using the Final Model for Forecasting. There are several excellent references to BJ methodology; see [14, 19] (which this presentation draws from), and the classic [11], which focuses on auto-regressive integrated moving average ($ARIMA$) models (a bit more general than the auto-regressive moving average models – ARMA – models treated here). For brevity, we omit discussions of 'invertibility', the roots of characteristic polynomials, state-space models and the Kalman filter for estimating the state vector and other complexities. Also, the models for Y_n in the classical BJ framework do not include any explanatory variables; this will be considered in a short subsection after the BJ methodology and its Bayesian analog are presented.

3.1. Model class identification

There are several model classes that recur throughout introductory time series. From a strictly operational level, the first task is to identify which one to use. We do not address this in general; we just look at the two most fundamental classes. These are the autoregressive process of order p, $AR(p)$, and the moving average process of order q, $MA(q)$. Given $\phi_1, \ldots, \phi_p$ the $AR(p)$ process is

$$Y_n = \nu + \phi_1 Y_{n-1} + \cdots + \phi_p Y_{n-p} + \epsilon_n,$$

where ν is a constant, here taken as zero unless otherwise noted. We assume that the roots of the polynomial

$$1 - \sum_{j=1}^{p} \phi_j z^j$$

satisfy $|z_i| > 1$ for $i = 1, \ldots, p$ to ensure stationarity. (For an $AR(1)$ process this is equivalent to $|\phi_1| < 1$). The $MA(q)$ process is defined for parameters $\theta_1, \ldots, \theta_q$ by

$$Y_n = \mu + \epsilon_n - \theta_1 \epsilon_{n-1} - \cdots - \theta_q \epsilon_{n-q}$$

where the ϵ_n's are IID mean zero, variance σ^2 and μ is the mean, here taken to be zero unles otherwise noted. We assume the roots of the polynomial

$$1 - \sum_{j=1}^{q} \theta_j x^j$$

are outside the unit circle in $\mathbb{C}$. Taken together, the $ARMA(p,q)$ process is

$$Y_n = c + \phi_1 Y_{n-1} + \cdots + \phi_p Y_{n-p} + \epsilon_n - \theta_1 \epsilon_{n-1} - \cdots - \theta_q \epsilon_{n-q}, \tag{3.1}$$

where c is a constant, again taken to be zero unless otherwise noted.

If it is known that a given time series is $MA(q)$, model identification means choosing q. If it is known that a given time series is $AR(p)$, model identification means choosing p. If it is known that a given time series is $ARMA(p,q)$, then the task is identifying p and q. More general classes have also been well-studied, but we do not present those results here.

The autocorrelation function (ACF) of a time series is $\rho(k) = \gamma(k)/\gamma(0)$ at lag k, where

$$\gamma(k) = \frac{E[(Y_n - \mu)(Y_{n+k} - \mu)]}{\sigma^2}$$

where Y_n has mean μ and variance σ^2. Clearly, $\gamma(-k) = \gamma(k)$ if both sides are defined. It can be verified that for an $MA(q)$ process, the ACF is zero for $k > q$. The partial autocorrelation of a time series is more complicated. Given a stationary time series Y_n we can write the regression function

$$E(Y_n | Y_{n-1} = y_{n-1}, \ldots, Y_{n-k} = y_{n-k}) = \beta_{k,1} y_{n-1} + \ldots + \beta_{k,k} y_{n-k}.$$

Each $\beta_{k,n-j}$ is the linear regression coefficient of Y_n on the $Y_{n-1}, \ldots, Y_{n-k}$ treated as explanatory variables. From standard linear regression theory, the last of these is

$$\beta_{k,k} = \mathrm{Corr}(Y_n, Y_{n-k} | Y_{n-1}, \ldots, Y_{n-k+1}),$$

the dependence between Y_n and Y_{n-k} that cannot be accounted for by the intervening $Y_{n-1}, \ldots, Y_{n-k+1}$. The sequence $\beta_{1,1}, \ldots, \beta_{k,k}, \ldots$ is the partial autocorrelation function (PACF). It can be verified that for an $AR(p)$ process, the PACF is zero for $k > p$.

Now, we can select q and p by looking at plots of the ACF and PACF respectively for $k = 1, 2, \ldots$. First, the usual estimate for the ACF is

$$\hat{\rho}(k) = \frac{\sum_{u=k+1}^{n} (y_{u-k} - \bar{y})(y_u - \bar{y})}{\sum_{u=1}^{n} (y_u - \bar{y})^2}.$$

Although the $\hat{\rho}(k)$'s are themselves correlated, the correlations are weak enough that in the limit of large n we get

$$\mathrm{Var}(\hat{\rho}(k)) \approx \frac{1}{n} \quad \text{and} \quad \forall j \neq k, \mathrm{Corr}(\hat{\rho}(j), \hat{\rho}(k)) \approx 0.$$

So, for an $MA(q)$ process, if we plot the ACF values over time, we can look for the value of k for which the ACF falls into $0 \pm 1.96/\sqrt{n}$ (and stays there) and take it as our estimate $\hat{q}$ of q.

Second, we use the PACF to choose p. More exactly, we end up using the ACF again because we can re-express the PACF values in terms of the ACF values. This can be done either via the Yule-Walker equations or via the Levinson-Durbin recursions which we do not show here. The net effect is that the $\hat{\rho}(k)$'s can be used to generate estimates $\hat{\beta}_{k,k}$, for the PACF values. It can be shown that for an $AR(p)$ process, $\hat{\beta}_{k,k}$ are mean zero and variance $1/n$. Parallel to finding $\hat{q}$, if we plot the PACF values over time, we can look for the value of k for which the PACF falls in the interval $0 \pm 1.96/\sqrt{n}$ (and stays there) and take it as our estimate $\hat{p}$ of p.

For $ARMA(p,q)$ processes we *cannot* combine these two techniques (ACF and PACF) to find $\hat{p}$ and $\hat{q}$. This is so because including both AR and MA terms causes the ACF and PACF to have geometrically decaying patterns that cannot be clearly associated with any specific order (p when $q > 0$ or q when $p > 0$). Instead, we use an information criterion such as AIC or BIC (see Section 3.3) to determine the proper choices for p and q; for a review see [38].

3.2. Estimating parameters

Given that we have chosen appropriate $\hat{p}$ and $\hat{q}$, the next task is to estimate the θ_j's in the case of an $MA(p)$ process, the ϕ_j's in the case of an $AR(p)$ process, or both in the case of an $ARMA(p,q)$ process. This can be done by the method of moments, i.e., equate sample moments to population moments and solve the resulting implicit equations for parameter estimates, but the estimators obtained are quite inefficient. (Actually, the method of moments estimators for $AR(p)$ processes based on $\hat{\rho}$ are not too bad: They can be derived by solving the Yule-Walker equations $\rho(k) = \phi_1\rho(k-1) + \cdots + \phi_p\rho(k-p)$ for $k \geq 1$ and substituting $\hat{\rho(k)}$ for $\rho(k)$ for $k = 1, \ldots, p$. The analogous procedure for $AR(q)$ processes is much worse.)

The two methods that are typically used to obtain estimators are maximum likelihood (MLE's) and least squares (LSE's). To use maximum likelihood, one must have a likelihood. However, the statements so far have only used the first two moments of Y_n. So, we must choose a likelihood to generate estimates. If the error terms are assumed to be IID $N(0, \sigma^2)$ then any random vector $(Y_1, \ldots, Y_n)^T$ can be regarded as following an n-dimensional normal distribution. Consequently, the likelihood can be maximized – if the unobserved initial values of $\epsilon_0, \ldots, \epsilon_{-q+1}$ are properly dealt with. A simple technique for this is used below with $MA(q)$ processes. The downside of this likelihood approach is that it requires careful checking that the normality assumption is reasonable. Nevertheless, this is the method used in many software packages.

The least squares approach does not require likelihood assumptions and works much the same for time series as for linear regression. For instance, consider an

$AR(1)$ model

$$\epsilon_n = Y_n - \nu - \phi_1 Y_{n-1},$$

noting that $\nu = (1-\phi_1)\mu$ where $\mu = E(Y_n)$ for any n. The sum of squared errors is

$$S(\mu, \phi) = \sum_{k=1}^{n} (Y_k - \nu - \phi_1 Y_{k-1})^2, \tag{3.2}$$

where the default $Y_0 = 0$ is often chosen. Solving $\partial S/\partial \nu = 0$ and $\partial S/\partial \phi_1 = 0$ gives solutions $\hat{\phi}_1$ and $\hat{\nu}$ that can be found in closed form. Usually, $\hat{\nu} \approx \bar{Y}$ and $\hat{\phi}_1 \approx \hat{\rho}(1)$, i.e., the results are very similar to those for straight line regression (and the method of moments based on $\hat{\rho}$ for $AR(1)$ processes). The parallel continues for higher order AR processes and, asymptotically in n, the results are similar to those of linear regression. In particular, t-tests can be used to test any hypothesis of the form $\mathcal{H}_{0,j} : \phi_j = 0$ for $j = 1, \ldots, p$.

The LSE's for an $MA(q)$ process involve an extra wrinkle. If $\mu = 0$, the $MA(1)$ model leads to a sum of squared errors of the form

$$S(\theta) = \sum_{k=1}^{n} (Y_k + \theta_1 \epsilon_{k-1})^2. \tag{3.3}$$

The problem is that the noise terms ϵ_k are unknown. One way around this is to set the white noise process at time zero to be its mean, i.e., set $\epsilon_0 = 0$, so that $\hat{\epsilon}_j = Y_j + \theta_1 \hat{\epsilon}_{j-1}$ for $j = 1, \ldots, n$. Using these in (3.3) gives a new expression, say $S^*(\theta)$, that can be optimized, but not in closed form. Some sort of numerical optimization procedure must be used. In this procedure, it is important to use several plausible values of $\epsilon_0, \ldots, \epsilon_{-q+1}$ to be sure the parameter estimates are not unduly sensitive.

The same procedure can be used for higher order MA models, but for an $MA(q)$ process, one must set $\epsilon_0 = \cdots = \epsilon_{-q+1} = 0$. It is seen that any process with a moving average component will have a problem with the initial ϵs and that any choice for ϵ_0 (for instance) will affect the solutions and propagate over time. Two ways around this are (i) sensitivity analysis: choose other values for $\epsilon_0, \ldots, \epsilon_{-q+1}$ and decide if they affect the parameter estimates overmuch, and (ii) refitting: Assume $\epsilon_0, \ldots, \epsilon_{-q+1}$ and get estimates of $(\theta_1, \ldots, \theta_q)$, use these with the Y_i's to find updated values of $\epsilon_0, \ldots, \epsilon_{-q+1}$ which can then be used again to generate new estimates for $(\theta_1, \ldots, \theta_q)$ cycling until stable values for $\epsilon_0, \ldots, \epsilon_{-q+1}$ are found. It turns out that usually the estimates of $(\theta_1, \ldots, \theta_q)$ are fairly stable unless n is small (or the model is nearly non-invertible, a case we do not cover here).

The methods for (3.2) and (3.3), and their extensions to general p and q, can be combined to give LSE's for general $ARMA$ processes. The mean zero $ARMA(1,1)$ process can be written as

$$\epsilon_n = Y_n - \phi_1 Y_{n-1} + \theta_1 \epsilon_{n-1},$$

leading to

$$S(\phi_1, \theta_1) = \sum_{k=1}^{n} (Y_k - \phi_1 Y_{k-1} + \theta_1 \epsilon_{k-1})^2.$$

Setting $Y_0 = \epsilon_0 = 0$ permits a numerical solution for $\hat{\phi}_1$ and $\hat{\theta}_1$ and the general case ($ARMA(p, q)$) would require $\epsilon_0 = \cdots = \epsilon_{-q+1} = 0$. As in the $MA(q)$ case, the effect of the initial value will propagate over time and suitable checks should be made.

3.3. Validation

There are two sorts of validation to be done: Verification that the correct p or q has been found (assuming an $ARMA$ model is correct in the first place) and given that p and q are correct, verification that the parameters in the model are properly estimated.

There are two standard ways to evaluate whether p and q are plausible. First, one can use several model selection techniques like the Akaike information criterion (AIC) or the Bayes information criterion (BIC) and compare the resulting choices for p and q. Recall that the AIC is

$$AIC(p, q) = -\ln L(\theta, \psi | y^n) + k$$

where $L(\theta, \psi | y^n)$ is the likelihood, $k = p + q$, and $\theta = (\theta_1, \ldots, \theta_q)$, $\phi = (\phi_1, \ldots, \phi_p)$. Finding the pair (p, q) that minimizes the AIC often provides a check of the $\hat{p}$ and $\hat{q}$ found using the ACF and PACF. BIC, like the AIC, relies on having a likelihood (usually normal), but has a heavier penalty because k is multiplied by the log sample size:

$$BIC(p, q) = -2 \ln L(\theta, \psi | y^n) + k \ln n.$$

Because of the larger penalty on the number of parameters, the BIC tends to select smaller models than the AIC. Both the AIC and BIC can be written in terms of the residual error, usually written as '$s^2 = SSE/df(SSE)$' in the normal error case. That is,

$$AIC = n \ln s^2 + 2k \quad \text{and} \quad BIC = n \ln s^2 + k \ln n.$$

There have been extensive comparisons of AIC versus BIC in a wide variety of contexts, see [15] (Chap. 6, Sec. 4) and [26] (Chap. 10, Sec. 2) for summaries and [92] and [42] for recent contributions. Roughly, AIC identifies a model, often depending heavily on n, that is good in a variance-bias tradeoff sense. This often enables models chosen by the AIC to give better predictions when the true model is hard to specify, e.g., in an $\mathcal{M}$-complete or $\mathcal{M}$-open setting, see [5]. (Loosely, $\mathcal{M}$-complete means that the true model is only approximable by models under consideration while $\mathcal{M}$-open means that the true model is even more inaccessible.)

By contrast, BIC identifies the model from those available that is closest to the true model in relative entropy. (AIC has a different relative entropy interpretation based on a prediction criterion.) This often enables models chosen by the BIC to give better predictions when the true model is in the model class under consideration i.e., $\mathcal{M}$-closed settings see [5], or when the model approximation error is much smaller than any other source of error. Overall, the BIC favors smaller models (given that they are equally good); the AIC tends to favor larger models provided they are at least a little better than any smaller model.

In practice the two criteria often give similar results – at least when the sample sizes are moderate, the model sizes are not too large, and the true model is not too far from the list of models under consideation. See [52] for further comparison of AIC, BIC, and the Hannan-Quinn criterion. Since there are numerous information criteria, if they give similar choices for p and q then one may regard the determination of p and q as more reliable.

Another technique for satisfying oneself that the $\hat{p}$ and $\hat{q}$ are reasonable is to overfit and prune back by testing. That is, if one method such as the ACF and PACF led to $\hat{p}$ and $\hat{q}$, one might look at the fit using an $ARMA(\hat{p}+1, \hat{q})$ or an $ARMA(\hat{p}, \hat{q}+1)$ and use a t-test to see if the extra parameter can be set to zero. This approach is logical and simple to implement but ignores the highly collinear nature of the AR terms. As in linear regression one can calculate $R^2 = 1 - s^2/s_y^2$ where s_y^2 is the variance of the observations, or modifications of it such as the adjusted R^2, to compare various models. A caveat to this approach is seen in [19] who note that in an $AR(1)$ model, the process variance is $\gamma(0) = \sigma^2/(1-\phi_1)^2$ and if we knew the correct model we would get $R^2 = 1 - \sigma^2(1-\phi_1)^2/\sigma^2 = 2\phi_1 - \phi_1^2$. Then, if $\phi = .2$ we would get $R^2 = .36$ – commonly regarded as small but here representing the best one could possibly do because of the intrinsic variability in the dependent process. It is also a fact that models with high values of R^2 do not necessarily provide good forecasts.

Given that a satisfactory choice of p and q has been made, the usual sort of residual analysis is done conditionally on the model selection, i.e., ignoring the variability in the $\hat{p}$ and $\hat{q}$, to ensure the parameter estimates are not too far wrong. That is, if the model class $(\hat{p}, \hat{q})$ is taken to be correct, the residuals $e_i = y_i - \hat{y}_i$ plotted over time should look IID mean zero and constant variance. Here, the y_i's are the data points and $\hat{y}_i$'s are the fitted values at time i using model $(\hat{p}, \hat{q})$ with parameter estimates $\hat{\theta}_1, \ldots, \hat{\theta}_q$ or $\hat{\phi}_1, \ldots, \hat{\phi}_p$, with the obvious simplifications if $p = 0$ or $q = 0$. Of course, one should check that the histogram of the residuals looks normal as well, particularly as normality of the residuals is important in constructing prediction intervals. Then, the main remaining task is to check the residuals are uncorrelated. This can be examined by plotting the sample autocorrelations

$$\hat{\rho}(k) = \frac{\sum_{i=1}^{n-k}(e_i - \bar{e})(e_{i+k} - \bar{e})}{\sum i = 1^n (e_i - \bar{e})^2}$$

as a function of k and making sure most of them land in the interval $\pm 2/\sqrt{n}$.

There are formal testing procedures for whether the sample autocorrelations are close enough to zero that they can be taken as zero. Many of these, like the Box-Pierce or Ljung-Box-Pierce statistics (i.e., portmanteau statistics) are frequentist and compare a sum of squared errors to percentiles from a Chi-squared distribution, effectively evaluating whether fitting the residuals with an IID model satisfies a goodness of fit criterion.

3.4. Forecasting

Given the successful completion of the first three stages, we can now turn to using the fitted model to generate predictions. So, suppose we have n observations $Y_1, \ldots, Y_n$ and we want to predict Y_{n+1}. One standard choice is

$$\hat{Y}_{n+1} = E(Y_{n+1}|Y_1, \ldots, Y_n) \tag{3.4}$$

the best approximation of Y_{n+1} with respect to squared error. It can be computed if the parameters are known; more typically, they are not known, so parameter estimates are just plugged in to $E(Y_{n+1}|Y^n)$ giving a predictor that is often not too bad. Thus, since the earlier subsections enable us to identify $\hat{Y}_{n+1}$ and, indeed, $\hat{Y}_{n+\ell}$ for ℓ step ahead predictions, we can generate point predictors readily. We also want a variance for our predictions, and the previous subsections enable us to give that, too.

For an $AR(1)$ process, (3.4) gives that

$$\hat{Y}_{n+\ell} = \nu(1 - \phi_1^\ell) + \phi_1^\ell Y_n \tag{3.5}$$

so that $\hat{Y}_{n+\ell} \to \nu \approx \bar{Y}$ as ℓ increases since $|\phi_1| < 1$. To get a point predictor for Y_{n+1} we substitute the estimates $\hat{\mu}$ and $\hat{\phi}_1$ and the observation y_n into (3.5).

For an $MA(1)$ process, (3.4) gives

$$\hat{Y}_{n+1} = \mu - \theta_1 E(\epsilon_n|Y_1, \ldots, Y_n) = \mu - \theta_1 \epsilon_n \approx \mu - \theta_1 e_n. \tag{3.6}$$

In (3.6) we have used $E(\epsilon_n|Y_1, \ldots, Y_n) \approx e_n$; this is not likely to be too far wrong when is it reasonable to regard the e_n's as IID outcomes of the noise distribution. (Ensuring this was part of the intent behind the residual analysis at the end of Sec. 3.3.) To get a point predictor we substitute the estimates $\hat{\mu}$ and $\hat{\theta}_1$ into (3.6). For $\ell > 1$ we can derive the simpler formula

$$\hat{Y}_{n+\ell} = \mu + E(\epsilon_{n+\ell}|Y_n, \ldots, Y_1) - \theta_1 E(\epsilon_{n+\ell-1}|Y_n, \ldots, Y_1) = \mu,$$

and substitute $\hat{\mu}$ into it to get point predictions for $Y_{n+\ell}$.

Putting these two simpler cases together we can give the expression for a mean zero $ARMA(p, q)$ process. We use $Y_n = \sum_{k=1}^{p} \phi_k Y_{n-k} + \epsilon_n - \sum_{j=1}^{q} \theta_j \epsilon_{n-j}$ in (3.4) to obtain

$$\hat{Y}_{n+\ell} = \sum_{k=1}^{p} \phi_k \hat{Y}_{n+\ell-k} - \sum_{j=1}^{q} \theta_j E(\epsilon_{n+\ell-j}|Y_n, \ldots, Y_1) \tag{3.7}$$

where we set $\hat{Y}_{n+j} = y_{n+j}$ for $j \leq 0$. The conditional expectations can be found using

$$E(\epsilon_{n+u}|Y_n, \ldots, Y_1) \approx \begin{cases} 0 & \text{if } u \geq 1, \\ e_{n+u} & \text{if } u \leq 0 \end{cases} \tag{3.8}$$

in which $u = \ell - j$ and $e_{n+u} = e_n(u) = y_{n+u} - \hat{y}_{n+u}$. Thus, plugging in estimates of θ_j, ϕ_k and using the y_i's gives point predictors iteratively starting with $\ell = 1$ to get $\hat{y}_{n+1}$ and moving on to $\ell = 2, 3, \ldots$.

Strictly speaking, under the Prequential Principle, to evaluate a predictor it is enough to have a sequence of obervations and their corresponding point predictions. However, predictive variances remain important, partially because they are required to give PI's. So, recall the $AR(1)$ process for which an ℓ step ahead point predictor was just given. The ℓ step ahead residual is $e_n(\ell) = y_{n+\ell} - \hat{y}_{n+\ell}$. One can derive $E(e_n(\ell)) = 0$ by using the $MA(\infty)$ representation of an $AR(1)$ model, namely

$$Y_n - \nu = \sum_{k=0}^{\infty} \phi^k \epsilon_{n-k},$$

which follows from the recursive definition of Y_n and the expression

$$\begin{aligned} \hat{Y}_{n+\ell} &= E(Y_{n+\ell}|Y_1^n) = \nu + \phi[E(Y_{n+\ell-1}|Y_1^n) - \nu] + E(\epsilon_{n+\ell}|Y_1^n) \\ &= \nu + \phi[\hat{Y}_{n+\ell-1} - \nu] \\ &= \ldots = \nu + \phi^\ell(Y_n - \nu), \end{aligned}$$

from ℓ uses of the definition of an $AR(1)$ process. Now, as a random variable, the 'predictual' is

$$\begin{aligned} e_n(\ell) &= Y_{n+\ell} - \nu - \phi^\ell(Y_n - \nu) \\ &= \sum_{k=0}^{\infty} \phi^k \epsilon_{n+\ell-k} - \phi^\ell \sum_{k=0}^{\infty} \phi^k \epsilon_{n-k} \\ &= \epsilon_{n+\ell} + \phi\epsilon_{n+\ell-1} + \cdots + \phi^{\ell-1}\epsilon_{n+1}. \end{aligned}$$

So, taking expectations on both sides gives $E(e_n(\ell)) = 0$ which means that $\hat{Y}_{n+\ell}$ is unbiased as a predictor for $Y_{n+\ell}$. Also, $\mathrm{Var}(e_n(\ell)) = \sigma^2(1-\phi_1^{2\ell})/(1-\phi_1^2)$ which tends to $\sigma^2/(1-\phi_1^2)$ the marginal variance $\mathrm{Var}(Y_n)$ of an $AR(1)$ process for any n, provided $|\phi_1| < 1$. A prediction interval is formed by plugging in estimates to give $\hat{Y}_{n+\ell} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(e_n(\ell))}$ as an approximate $(1-\alpha)$ PI.

Forming a PI for an $MA(1)$ process is a little more complicated. However, as with an $AR(1)$ process, it can be shown that for $MA(1)$ processes that $E(e_n(\ell)) = 0$ and hence $\hat{Y}_{n+\ell} = \mu$. So, it is reasonable to get the slightly modified prediction $\hat{Y}_{n+\ell} = \hat{\mu}$ since $\hat{\mu}$ is unbiased for μ. In addition, $\mathrm{Var}(e_n(1)) = \sigma^2$ and for $\ell > 1$, $\mathrm{Var}(e_n(\ell)) = \mathrm{Var}(Y) = \sigma^2(1+\theta_1^2)$. Again, PIs can be found from

$\hat{Y}_{n+\ell} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(e_n(\ell))}$ for $\ell \geq 1$. This discussion assumes that we know ϵ_n, and in turn ϵ_{n-1}, ϵ_{n-2} and so forth down to ϵ_0. For a large class of AR models, however, the approximation $E(\epsilon_n|Y_1, \ldots, Y_n) \approx e_n$ will be valid when n is large. (The required condition is called invertibility which we do not discuss here.) We can also use $e_1, \ldots, e_{n-1}$, the previous errors, as predictions of $\epsilon_1, \ldots, \epsilon_{n-1}$.

In the case of a stationary $ARMA(p, q)$ model, parallel to the $AR(1)$ process, it can be shown that $E(e_n(\ell)) = 0$ and $\mathrm{Var}(e_n(\ell)) = \sigma^2 \sum_{j=0}^{\ell-1} \psi_j^2$, where the ψ_j emerge from the $MA(\infty)$ representation of an $ARMA$ model, namely that

$$Y_n - c = \sum_{k=0}^{\infty} \psi_k \epsilon_{n-k},$$

for some sequence of weights ψ_k. For the special case of $p = q = 1$, we can show that $Y_{n+\ell} = c + \phi_1(Y_{n+\ell-1} - \mu) + \epsilon_{n+\ell} - \theta_1\epsilon_{n+\ell-1}$ and therefore our predictor is

$$\hat{Y}_{n+\ell} = \hat{\mu} + \hat{\phi}_1(\hat{Y}_{n+\ell-1} - \hat{\mu}) + \hat{\epsilon}_{n+\ell} - \hat{\theta}_1\hat{\epsilon}_{n+\ell-1}.$$

In this expression, $\hat{\mu}$ is the sample mean (estimating c) and $\hat{\phi}_1$ and $\hat{\theta}_1$ are LSE's (or MLE's if a likelihood can be found justified). The $\hat{\epsilon}$'s are the fitted values of the corresponding ϵ's from the least squares estimation in the $MA(1)$ component of the model; more generally, in this expression, 'hat' over a random variable indicates its conditional expectation given $Y_1^n = (Y_1, \ldots, Y_n)$ and 'hat' over a parameter indicates an estimate of the parameter using $y_1^n = (y_1, \ldots, y_n)$.

When $\ell = 1$, this is $\hat{Y}_{n+1} = \hat{\mu} + \hat{\phi}_1(Y_n - \hat{\mu}) - \hat{\theta}_1 e_n$ and this generalizes to $\ell \geq 2$: $\hat{Y}_{n+\ell} = \hat{\mu} + \hat{\phi}_1^{\ell}(Y_n - \hat{\mu}) - \hat{\phi}_1^{\ell-1}\hat{\theta}_1 e_n$. Using the $MA(\infty)$ representation of an $ARMA$ model, it can be shown that $\psi_0 = 1$ and $\psi_j = \phi_1^{j-1}(\phi_1 - \theta_1)$ for $j \geq 1$, see [19], so that in addition to $E(e_n(\ell)) = 0$ we have

$$\mathrm{Var}(e_n(\ell)) = \sigma^2 \left(1 + \frac{(\phi_1 - \theta_1)^2(1 - \phi_1^{2(\ell-1)})}{1 - \phi_1^2}\right),$$

that goes to $\sigma^2(1 + (\phi_1 - \theta_1)^2/(1 - \phi_1^2))$, the unconditional variance of the Y_i's. So, since we can estimate the θ_1 and ϕ_1 we can form prediction intervals $\hat{Y}_{n+\ell} \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(e_n(\ell))}$. Other values of p and q can be done, but are more complicated.

3.5. Bayesian approach for ARMA models

The Bayesian approach to time series begins with [100] and [98]; [12] also devotes a section to the Bayesian treatment. Here, we look at the predictors found from Bayesian analysis of $ARMA$ models beginning with the first complete analysis given in [73]. Then we look at some aspects of robustness and the selection of the p and q in the $ARMA$ model. In the Appendix, we look at the dynamic linear model which is a generalization of the classical models.

To present the method of [73], write the $ARMA(p,q)$ process as

$$(Y_n-\mu)-\phi_1(Y_{n-1}-\mu)-\cdots-\phi_p(Y_{n-p}-\mu)=\epsilon_n-\theta_1\epsilon_{n-1}-\cdots-\theta_q\epsilon_{n-q}$$

in which the ϵ_n are IID $N(0,\sigma^2)$ and $n=-\infty,\ldots,-1,0,1,\ldots,\infty$. We include the centering explicitly since the mean $\mu=E(Y_n)$ is one of the parameters we want to estimate. We assume a finite string $Y_1^n=(Y_1,\ldots,Y_n)$ is available; it can be shown that $Y_1^n\sim MVN(\mu\mathbf{1}_n,\sigma^2A_n)$ where A_n is the $n\times n$ matrix with (i,j) elements given by $a_{i,j}=\mathrm{Cov}(Y_i,Y_j)=\rho(|i-j|)$. It is seen that ρ is a function of $\phi_1,\ldots,\phi_p$ and $\theta_1,\ldots,\theta_q$.

Now suppose we want to predict $Y_{n+1}^{n+\ell}=(Y_{n+1},\ldots,Y_{n+\ell})^T$, i.e., forecast ℓ steps ahead. We have a joint normal distribution for $(Y_1,\ldots,Y_{n+\ell})$ with variance matrix $A_{n+\ell}$ that can be partitioned into blocks. Write it as

$$A_{n+\ell}=\begin{pmatrix} A_n & A_{12}\\ A_{21} & A_\ell\end{pmatrix} \tag{3.9}$$

and set $A_{n:\ell}=A_\ell-A_{21}A_n^{-1}A_{12}$. Now, standard normal theory gives that $(Y_{n+1}^{n+\ell}|Y_1^n)$ is an ℓ-dimensional normal with mean $\mu\mathbf{1}_\ell+A_{21}A_n^{-1}(Y_1^n-\mu\mathbf{1}_n)$ and covariance matrix $A_{n:\ell}$. Provided the time series is stationary ($1-\sum\phi_iz^i=0$ has roots outside the unit circle in $\mathbb{C}$) and invertible (the roots of $1-\sum\theta_iz^i=0$ lie on or outside the unit circle in $\mathbb{C}$ allowing an MA process to be represented as an infinite AR process) there will be unique values μ, σ, p, $\phi_1,\ldots,\phi_p$, q and $\theta_1,\ldots,\theta_q$ that identify the true model.

Given unique parametrizations, the remaining tasks in a full Bayes analysis are to specify the prior distribution for $(p,q,\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q,\mu,\sigma^2)$ and to obtain the posterior. To begin the prior specification, order the possible pairs (p,q) by dictionary order on p for each fixed value of $p+q$. Thus, for $p+q=0$, there is only one possibility, $(0,0)$. For $p+q=1$, there are two possibilities and the dictionary order is $(0,1)$ and $(1,0)$. For $p+q=2$, dictionary order gives $(0,2)$, $(1,1)$, and $(2,0)$ and so forth. In principle, any probability mass function with support equal to all pairs of non-negative integers will permit inferences; choose one and denote it generically by $w(p,q)$.

Conditional on the choice of p and q, we must specify a prior $w(\phi_1^p,\theta_1^q|p,q)$ on a subset of the ϕ_1^p's and θ_1^q's for which the process $\langle Y_n\rangle$ is stationary and invertible. The constraint on the values of ϕ_1^p and θ_1^q are not regarded here as part of the prior information; they are necessary for identifiability of a stable model. Given a choice for $w(\phi_1^p,\theta_1^q|p,q)$ it remains to select priors for μ and $r=1/\sigma^2$ conditional on p, q, ϕ_1^p, and θ_1^q. One standard choice is

$$w(r|p,q,\phi_1^p,\theta_1^q)\sim\mathrm{Gamma}(\alpha,\beta)\quad\text{and}\quad w(\mu|r,p,q,\phi_1^p,\theta_1^q)\sim N(\gamma,\frac{1}{\tau r})$$

in which α, β, γ, and τ are hyperparameters.

Since we have that

$$(Y_1^n|p,q,\phi_1^p,\theta_1^q,\mu,r)\sim N(\mu\mathbf{1}_n,A_n/r),$$

one can integrate over μ and r to obtain a multivariate t distribution. This is identified in [73] as

$$(Y_1^n|p,q,\phi_1^p,\theta_1^q) \sim t_{n,2\alpha}(\gamma \mathbf{1}_n, (\alpha/\beta)[A_n + \mathbf{1}_n\mathbf{1}_n^T/\tau]^{-1}),$$

denoted $p(y_1^n|p,q,\phi_1^p,\theta_1^q)$, where the arguments of the t are the location and scaling and the subscripts are degrees of freedom and the hyperparatmer α. Similar arguments on the conditional of $Y_{n+1}^{n+\ell}$ given Y_1^n result in

$$\begin{aligned}(Y_{n+1}^{n+\ell}|Y_1^n = y_1^n, p, q, \phi_1^p, \theta_1^q)\\ \sim t_{n,2\alpha}(\gamma^*\mathbf{1}_n + A_{21}A_n^{-1}(y_1^n - \mu\mathbf{1}_n), (2\alpha + n/2\beta^*)[A^*_{n:\ell} + \tau^{-*}\mathbf{1}_\ell\mathbf{1}_\ell^T]^{-1}),\end{aligned}$$

denoted $p(Y_{n+1}^{n+\ell}|y_1^n, \phi_1^p, \theta_1^q, p, q)$ and where

$$\begin{aligned}\gamma^* &= (\gamma\tau + \mathbf{1}_\ell^T A_n^{-1}\mathbf{1}_\ell)/\tau^*,\\ \tau^* &= 1/\tau^{-*} = \tau + \mathbf{1}_n^T A_n^{-1}\mathbf{1}_n,\end{aligned}$$

and

$$\beta^* = \beta + (y_1^n - \gamma\mathbf{1}_n)^T(A_n + \mathbf{1}_n\mathbf{1}_n^T/\tau)^{-1}(y_1^n - \gamma\mathbf{1}_n);$$

see [73].

Now, the Bayesian forecaster who believes the relative entropy loss is relevant uses the predictive density (2.11); this is an easy extension of the optimality of (1.2) established in [2]. The following distributions were derived in [72]. The marginal of $Y_1^{n+\ell}$ is

$$p(y_1^n|p,q) = \int w(\phi_1^p, \theta_1^q|p,q)p(y_1^n|\phi_1^p,\theta_1^q,p,q)d\phi_1^p d\theta_1^q.$$

So, Bayes' rule gives

$$w(\phi_1^p, \theta_1^q|p,q,y_1^n) = \frac{w(\phi_1^p,\theta_1^q|p,q)p(y_1^n|\phi_1^p,\theta_1^q,p,q)}{p(y_1^n|p,q)},$$

and

$$w(p,q|y_1^n) = \frac{w(p,q)p(y_1^n|p,q)}{\sum_{p,q} w(p,q)p(y_1^n|p,q)}$$

and finally the predictive density $p(y_{n+1}^{n+\ell}|y_1^n)$ equals

$$\sum_{p,q} w(p,q|y_1^n) \int p(y_{n+1}^{n+\ell}|y_1^n,\phi_1^p,\theta_1^q,p,q)w(\phi_1^p,\theta_1^q|y_1^n,p,q)d\phi_1^p d\theta_1^q. \tag{3.10}$$

Expression (3.10) is an average over models and the uncertainty in p and q automatically affects the width of the prediction intervals. Computational approaches to evaluating these numerically are given in [73], but Bayesian computing has advanced much beyond them. So, contemporary techniques such as MCMC are more typically used but not discussed here; see [78] for one example.

A key problem with any time series analysis is that the model fitting, i.e., estimating the parameters, is confounded with the definition of outliers. That is, a given data point may be an outlier for one fitted model, but not for another. Consequently, identification of outliers or other influential data points is dependent on the parameter estimates that they influence. This problem is taken up in [6] who propose a robust version of the Monahan procedure above based on writing $Z_n = Y_n + O_n$ where Y_n is the usual $ARMA(p,q)$ series but only Z_n is observed. The additive term O_n is an outlier. An added feature of the set up in [6] is that there is positive prior probability on some of the θ's and ϕ's being zero. The outliers O_t and the ϵ's in Y_t are permitted to have distributions given as finite mixtures of normals so while the tails are not heavier, the effective range is larger. A weaker form of robustness limited to the prior was examined in [101], who used different priors representing different sorts of economic assumptions to evaluate the effect on inferences.

3.6. Explanatory variables

Here we look at the connection between $ARMA$ models and linear regression. This arises in two ways. The simplest is that the error term in a linear regression is $ARMA(p,q)$ rather than IID. In more complex settings, the time series structure may extend to the explanatory variables as well, leading to the dynamic linear model (DLM). This last case, while fascinating, is quite complex and will only be explained cursorily in Appendix A.

3.6.1. $ARMA(p,q)$ error term

A linear model with $ARMA(p,q)$ error term can be given either a Bayesian or Frequentist analysis. Both analyses start with

$$Y = X\beta + U \tag{3.11}$$

where $Y = (Y_1, \ldots, Y_n)^T$, X is an $n \times k$ matrix, β is a $k \times 1$ parameter vector and $U = (U_1, \ldots, U_n)^T$ is an $n \times 1$ vector of random outcomes from a stationary $ARMA(p,q)$ process as in (3.1). That is, replace the Y_i's in (3.1) with the U_i's in (3.11). Thus, each U_i is a linear combination of $\langle \epsilon_j \rangle_{-\infty}^{i} = (\epsilon_{-\infty}, \ldots, \epsilon_i)$ (using the $MA(\infty)$ representation of the AR model). It is assumed that the roots of the AR part are outside the unit circle in $\mathbb{C}$ (stationarity) and all the roots of the MA part are on or outside the unit circle $\mathbb{C}$ (invertibility).

From the Frequentist perspective, two estimation procedures for the parameters have been studied in [103]. One is an ML approach and the other is a two stage procedure that estimates the $ARMA$ parameters first and then uses a generalized least squares approach to estimate β. Here we present only their ML procedure. To do this, we must define a variance matrix for a finite string of variables in a doubly infinite stochastic vector. So, recognize that for a doubly infinite vector $\langle U_i \rangle|_{-\infty}^{\infty}$ the covariance matrix is $\Sigma_\infty = E(UU^T)$ and we can

write $\langle U_i\rangle|_{-\infty}^{\infty} = \Sigma_{\infty}^{1/2}\langle\epsilon_i\rangle|_{-\infty}^{\infty}$. Since we have sample size n, we want to extract an $n \times n$ block of Σ_∞ to represent the variance matrix of a string of U_i's. It is enough to extract this as a block on the main diagonal of Σ_∞. This can be done using a doubly infinite projection matrix π_n given by the identity on the block we want to extract and all other entries zero. Now, for a finite string $U_1, \ldots, U_n$ of the doubly infinite stochastic process, we have $\Sigma_n = \pi_n \Sigma_\infty \pi_n$.

Now, if $\langle\epsilon_i\rangle|_{-\infty}^{\infty}$ has IID $N(0, \sigma^2)$ elements, the log-likelihood function is

$$\log L = const - \frac{1}{2}\log(\det \Sigma_n) - \frac{1}{2}u^T \Sigma_n^{-1} u,$$

where *const* is a constant. For fixed p and q and $\tau = (\tau_1, \ldots, \tau_{p+q}) = (\phi_1^p, \theta_1^q)$ the likelihood equations are

$$\frac{\partial \log L}{\partial \beta_k} = 0 \quad \text{and} \quad \frac{\partial \log L}{\partial \tau_m} = 0.$$

[103] shows these are equivalent to

$$\begin{aligned} \hat{\beta}_{ML} &= (X^T \Sigma_n X)^{-1} X^T \Sigma_n^{-1} y, \\ -\frac{1}{\det \Sigma_n}\left[\frac{\partial \det \Sigma_n}{\partial \tau_m}\right] &= u_{ML}^T \frac{\partial \Sigma_n^{-1}}{\partial \tau_m} u_{ML}^T \quad m = 1, \ldots, p+q \\ u_{ML} &= y - X\hat{\beta}_{ML}. \end{aligned} \tag{3.12}$$

For many such systems of equations, the estimates $\hat{\beta}_{ML}$ are $\sqrt{n}$ consistent, i.e., $\sqrt{n}(\hat{\beta}_{ML} - \beta)$ has a nontrivial limit. When this holds, the system (3.12) is essentially a set of solvable (polynomial) constraints on the parameters β and τ. Convergence of linear estimators such as these is in mean square and hence in probability. In fact, [103] shows that a simplified form of these estimators using method of moments reasoning converges as well.

As in the earlier cases, a point prediction for a new input vector ℓ steps into the future, $X_{n+\ell}$, would be given by $\hat{Y}(X_{n+\ell}) = X_{n+\ell}\hat{\beta}_{ML}$ with prediction intervals coming from $\pm z_{1-\alpha/2}\sqrt{\widehat{Var}(e_n(\ell)}$ as in Sec. 3.4.

Least squares is merely one choice of loss function among many that are possible. For instance, least absolute deviation estimators have been studied in the $ARMA(p, q)$ error context and [34] has established consistency and asymptotic normality for them.

The Bayesian analysis that parallels this is given in [23]. They write

$$Y_i = X_i^T \beta + \epsilon_i \tag{3.13}$$

for $i = 1, \ldots, n$ where the ϵ_i's follow an $ARMA(p, q)$ as in (3.1). That is, the role of Y_n in (3.1) is played by ϵ_n and the role of ϵ_n is played by a perturbation term, say u_n, i.e.,

$$\epsilon_i = \phi_1\epsilon_{i-1} + \cdots + u_i + \theta_1 u_{i-1} + \cdots + \theta_q u_{i-q}.$$

It is seen that the error terms are dependent sequentially. The analysis in [23] rests on the use of a state space form described in [18], Chapter 5.5, but adapted to a linear model. ([23] does not give the details of this but cites an earlier text for the result.) For the normal error case, [23] assumes that p and q are known and that the roots of the characteristic polynomials for the AR and MA terms are outside the unit circle. Even in this case, the likelihood cannot be written down in a closed, tractable form. However, an expression for it can be obtained by using the conditional distributions of the predictors. Specifically,

$$p(y_i|\beta,\phi,\theta,\sigma^2;\epsilon_0,\ldots,\epsilon_{-p+1},u_0,\ldots,u_{-q+1}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-\hat{y}_{i|i-1})^2/2\sigma^2}$$

where $\hat{y}_{i|i-1}$ is the one-step-ahead prediction of Y_i given the information up to and including step $i-1$. An explicit form is given in [23]. Essentially, the n dimensional normal is factored into a sequence of conditional distributions analogous to that in (3.9) and the discussion following it. Using this, standard priors can be assigned to β, ϕ, θ and σ^2 and conditional posterior distributions can be derived. An MCMC algorithm can be designed to generate parameter estimates. Unfortunately, [23] does not give the forms of the predictions that would be obtained. However, the $\hat{\beta}$ produced by their method can be used in (3.13) to get point predictions. Prediction intervals can in principle be obtained from the predictive distribution using the technique described in Sec. 3.5.

4. Longitudinal

The key feature of longitudinal data is that many subjects are measured repeatedly, usually over time; sometimes this is called repeated measures data. This is opposed to cross-sectional data where individuals are measured once (usually at a fixed point in time). Longitudinal data differs from time series data because we have a population of individuals from which we have in general n samples whereas in time series it's as if $n = 1$, i.e., we have one sequence of data. In time series, it was the dependence structure that was the focus of attention. In this section, it will usually be the main effects that are of most interest.

One of the earliest efforts to focus on prediction in a longitudinal context is [81], see also the references therein. A recent text on longitudinal analysis that includes a (brief) treatment of prediction is [59], see also [45] and [94] (Chap. 3) that treats the linear mixed model as a 2-stage (or hierarchical) experiment.

Here, we present longitudinal analysis from a predictive standpoint. Having seen prediction in fixed effects linear models in Sec. 2.2.1 and Bayes linear models in Sec. 2.2.2, we turn to prediction in linear models that combine fixed and random effects which can be regarded mathematically as a sort of generalization of the Bayes structure. Then we turn to generalized linear mixed effects models and briefly to non-linear mixed effects models. This is done mostly for continuous responses. However, categorical longitudinal data can be analyzed by a wide variety of methods, even though we do not cover them; see [33] Chap. 7 and [1] Chap. 11. However, their perspective is model fitting not prediction.

4.1. Linear mixed models

We begin to think about longitudinal data by considering the linear mixed model for a single n_i-dimensional observation Y_i on subject i. So, write

$$Y_i = X_i\beta + Z_iU_i + \epsilon_i \tag{4.1}$$

for $i = 1, \ldots, n$ where $\beta = (\beta_1, \ldots, \beta_p)^T$ is the fixed effect (FE) with $n_i \times p$ design matrix X_i , $U_i \sim N_q(0, D)$ is the random effect (RE) with $n_i \times q$ design matrix Z_i and the error term is $\epsilon_i \sim N_{n_i}(0, R_i)$. The random effects are assumed IID and the error terms are mutually independent and independent of the U_i's. Here, subscripts on the normal distributions indicate the dimension; these are omitted when there will be no misunderstanding.

Model (4.1) can be regarded hierarchically. It is not hard to write down the within-subject version of (4.1) – basically (4.1) but setting $U_i = u_i$ – and separately write down the subject-to-subject variability defined by the distribution of U_i. The model (4.1) can also be regarded as a hierarchical Bayesian model in which $(Y_i|\beta, U_i, \theta)$ is specified first, where $D = D(\theta)$, and $(U_i|\theta)$ is specified second. The Bayesian model would be completed by specifying a prior for (β, θ). For brevity, we focus on the Frequentist story even though the Bayesian formulation may be useful in motivating computational techniques beyond our present scope. Whether one adopts a Bayes or Frequentist view, the inclusion of RE terms provides a lot more flexibility than fixed effect terms alone. Consequently, mixed effects (ME) models like (4.1) are often better able to summarize the information in the data more accurately.

The model (4.1) is used for each of n subjects so there are n versions of it i.e., n matrices X_i, and Z_i and n random variables U_i. The main way the different Y_i's are related to each other is by the FE component, specifically, β. Another way to think of this is that a longitudinal model is a collection of repeated time series so we can pool the data to estimate common features. So, for subject i, if we think of $j = 1, \ldots, n_i$ as time, it is sometimes better to write

$$Y_{ij} = \sum_{k=1}^{p} X_i(j;k)\beta_k + \sum_{k=1}^{q} Z_i(j;k)U_{ik} + \epsilon_i(j) \tag{4.2}$$

where $X_i(j;k)$ is the k-th element of the j-th row of X_i, i.e., the k-th explanatory variable measured at time j on subject i. The $Z_i(j;k)$'s are similar and $\epsilon_i(j)$ is the j-th element of ϵ_i.

4.1.1. Features of the model

Let us examine the meaning of (4.1) or (4.2). First, taking the expectation on both sides of (4.1) gives

$$EY_i = X_i\beta \quad \text{and} \quad EY_{ij} = \sum_{k=1}^{p} X_i(j;k)\beta_k$$

meaning that the random effects only matter at the subject level, not the population level. Otherwise put, individuals with high values of Z_iU_i are balanced by those with low values of Z_iU_i.

Next, we look at the two conditional expectations, Y on U and U on Y. The first is simpler. For $k \neq i$, $E(Y_k|U_i) = X_k\beta$ since U_i and Y_k are independent. However, when $i = k$ we get

$$E(Y_i|U_i) = X_i\beta + Z_iU_i, \tag{4.3}$$

because only the measurement error washes out. Both sides of expression (4.3) are random: Individual variability is represented as U_i and the term Z_iU_i represents the difference of individual i from the overall population mean $X_i\beta$. It is also easy to see that

$$\mathrm{Cov}(Y_i|U_i) = R_i.$$

For the second, we derive a closed form expression for $E(U_i|Y_i)$. For given i, we see that $Y_i \sim N_{n_i}(X_i\beta, V_i)$ where $V_i = \mathrm{Var}(Y_i) = Z_iDZ_i^T + R_i$. So, if all the n_i's are equal and the ϵ_i's are identical we can write $Y_i \sim N(X_i\beta, V_i)$ where $V_i = \mathrm{Var}(Y_i) = Z_iDZ_i^T + R$. A convenient simplification is to assume $R = \sigma^2 I$ where I is the identity matrix of dimension equal to the common value of the n_i's; this corresponds to the $\epsilon_{i,j}$'s being independent. It is easy to see that the covariance between Y_i and U_i is DZ_i^T. Thus, Y_i and U_i are jointly normally distributed, with dimension $n_i + q$, mean vector $(X_i\beta, \mathbf{0}_q)$ (where $\mathbf{0}_q$ is a vector of zero's of length q), and block covariance matrix as in

$$\begin{pmatrix} Y_i \\ U_i \end{pmatrix} \sim N\left(\begin{pmatrix} X_i\beta \\ \mathbf{0}_q \end{pmatrix}, \begin{pmatrix} V_i & Z_iD^T \\ DZ_i^T & D \end{pmatrix} \right).$$

Now, standard multivariate normal theory gives that

$$\begin{aligned} (U_i|Y_i) &\sim N_q(E(U_i|Y_i), \mathrm{Cov}(U_i|Y_i)) \\ &= N_q(DZ_i^T V_i^{-1}(Y_i - X_i\beta), D - DZ_i^T V_i^{-1} Z_i D). \end{aligned} \tag{4.4}$$

So, (4.4) identifies a closed-form expression for $E(U_i|Y_i)$. It is unrealistic to simplify (4.4) by setting $D = \sigma^2_{RE} I_q$ because then the Y_{ij} are independent even for fixed i. However, sometimes U_i is a scalar, $U_i \sim N(0, \sigma_{RE})$.

The best linear unbiased predictor (BLUP) for U_i is $E(U_i|Y_i)$ (in the usual squared error sense) using the generalized least squares estimate $\hat{\beta}_{GLS}$ for β in the first entry in (4.4). That is, let $\hat{U}_i = DZ_i^T V_i^{-1}(Y_i - X_i\hat{\beta}_{GLS})$. It is easy to see $E(U_i - E(U_i|Y_i)) = E(U_i - \hat{U}_i) = 0$ but harder to verify that the variance of $U_i - \hat{U}_i$ or $E(U_i|Y_i) - \hat{U}_i$ is minimal, even with the normality assumptions we have made. See [85] for a thorough treatment, with original references. However, the bigger problem is that we can not use (4.4) because it requires D, V_i and β be known and usually they are not. In the rest of this subsection we derive estimators for them; this is important to find models from which to generate predictions. We start with β and V_i.

First let us re-express (4.1) more concisely. Write

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \; X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \; Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}, \; U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}, \; \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{4.5}$$

and therefore

$$Y = X\beta + ZU + \epsilon,$$

leading to

$$Y \sim N(X\beta, V) \quad \text{where} \quad V = R + ZD_nZ^T$$

in which $R = \text{diag}(R_1, \ldots, R_n)$ and $D_n = \text{diag}(D, \ldots, D)$ i.e., n copies of D. (This suggests that a generalization to subject specific D is possible but we do not pursue it here.)

Now, for an individual i, the log-likelihood for Y_i is

$$\ell(V_i, \beta; y_i) = -\frac{1}{2}\left[\log|V_i| + (y_i - X_i\beta)V_i^{-1}(y_i - X_i\beta) + n_i \log(2\pi)\right]. \tag{4.6}$$

So, a maximum likelihood estimator (MLE) of β is $\hat{\beta}_i = (X_i^T V_i^{-1} X_i)^{-1} X_i^T V_i^{-1} Y_i$, if V_i is known. However, we can get higher efficiency by pooling the data over the n subjects since β is common to all subjects. The log-likelihood for Y is

$$\begin{aligned} \ell(V, \beta; y) = \quad & - \quad \frac{1}{2}\left[\log|V| + (y - X\beta)V^{-1}(y - X\beta) + n_i \log(2\pi)\right] \\ & + \quad \log(2\pi)\sum_{i=1}^{n} n_i. \end{aligned} \tag{4.7}$$

This gives that the MLE for β using all the data is

$$\hat{\beta} = \hat{\beta}_n = (X^T V^{-1} X)^{-1} X^T V^{-1} y = \left(\sum_{i=1}^{n} X_i^T V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{n} X_i^T V_i^{-1} Y_i, \tag{4.8}$$

provided V is known.

An analogous procedure can be used to get 'estimates' $\tilde{u}$– actually predictions – for the U_is (as well as $\tilde{\beta}$ for β). The joint distribution for U and ϵ is

$$\propto \begin{vmatrix} D_n & 0 \\ 0 & R \end{vmatrix}^{-1/2} \exp\left[-\frac{1}{2}\begin{bmatrix} u \\ y - X\beta - Zu \end{bmatrix}^T \begin{bmatrix} D_n^{-1} & 0 \\ 0 & R^{-1} \end{bmatrix} \begin{bmatrix} u \\ y - X\beta - Zu \end{bmatrix}\right].$$

Maximizing this over β and U follows by minimizing the negative of the exponent. Taking derivatives with respect to the components of β and U, setting them equal to zero and solving leads to Henderson's equations:

$$\begin{aligned} \begin{pmatrix} \tilde{\beta} \\ \tilde{u} \end{pmatrix} &= \begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + D_n^{-1} \end{pmatrix} \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix} \\ &= \begin{pmatrix} (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ D_n Z^T V^{-1}\left(y - X(X^T V^{-1} X)^{-1} X^T V^{-1} y\right) \end{pmatrix}. \end{aligned} \tag{4.9}$$

From (4.9) we see $\tilde{\beta}$ is what we already derived as the pooled MLE. Since the right hand term in parentheses in the lower entry is $\tilde{\beta}$, we get that $\tilde{u} = D_n Z^T V^{-1}(y - X\tilde{\beta})$ generalizing (4.4) from a a single i to all n subjects. However, this derivation does not show $\tilde{\beta}$ and $\tilde{u}$ are MLE's because the estimates are exhibited as a result of maximizing a joint distribution, not a true likelihood. However, the argument leading to (4.8) does mean that (4.8) is an MLE and the extension of (4.4) from U_i to U gives that $\tilde{U}$ is the BLUP (when V is known).

To get an estimate of β, it remains to estimate the V_i's for use in any of (4.4), (4.6), (4.7) i.e., (4.8), and (4.9). First recall that there are $\sum_i n_i$ measurements and that the V_i's represent a total of $\sum_i n_i(n_i - 1)$ values. If all $n_i = q$ then we have nq data points and $nq(q-1)$ parameters in the V_i's, an impossible situation. So, we must impose constraints on the V_i's for estimation to be feasible. For this reason, many authors write $R = R(\theta)$ where θ is the vector of components in R. This is particularly useful when it is permissible to assume $R = \sigma^2 I$ for some $\sigma > 0$. In this case we get $V_i = V_i(\theta, D)$ and $V = V(\theta, D)$ by setting $\theta = \sigma$. More compactly, we can write $V = V(\theta)$ by incorporating D into the components of θ, see [59]. Now, the estimate $\hat{\beta}$ from (4.8) formed by initially setting all $V_i = I_{n_i}$ (say) can be put into (4.7) for β to give the profile likelihood $\ell(V(\theta), \hat{\beta}; y)$ which can be maximized to give the MLE $\hat{V}$. This $\hat{V}$ can be put back into (4.8) to find a new $\hat{\beta}$ and one can cycle until convergence the estimates of V and β converge. (The MLE approach can also be used to find estimates for D and R_i directly, i.e., without estimating the V_i's; this is discussed briefly after 'REML' next but only for D. A Bayesian formulation is given in [66].)

An alternative way to find an estimate of V_i is called restricted maximum likelihood (REML). REML is often better than the MLE because the latter may have unacceptably high bias in small sample sizes. The basic idea, see [41] Chap. 4, is to find a $\sum_i n_i \times \sum_i n_i$ matrix K so that $E(KY) = 0$, then find the log-likelihood of $Y^* = (Y_1^{*T}, \ldots, Y_n^{*T})^T = KY$, and maximize it to find an estimates for the V_is. For instance, if K is block diagonal with blocks $K_i = I_{n_i} - X_i^T(X_i^T X_i)^{-1} X_i$ then $E(Y_i^*) = E(Y_i - X_i\hat{\beta}_i) = 0$ where $\hat{\beta}_i$ is the estimate of β formed by taking $V_i = I_{n+i}$. Now, Y_i^* has a multivariate normal distribution $N(0, K_i V_i K_i^T)$. Putting the Y_i^*'s together in the single vector Y^* and recalling $V = V(\theta)$, the log-likelihood given $Y^* = y^*$ is

$$\ell_R(\theta; y^*) = -\frac{1}{2}\left[\log|KVK^T| + y^{*T}(KV^{-1}K^T)^{-1}y^* + C\right], \tag{4.10}$$

where C is a constant independent of the parameters. Expression (4.10) can be differentiated with respect to θ and the derivatives set equal to zero to give a set of equations from which $\hat{V}$ and hence the $\hat{V}_i$'s can be found. Then V can be fed back into the definition of K (by the generalized least squares expression for the LSE using $\hat{V}$ rather than the identity matrix) to yield a new Y^* and a new form of (4.10). So, the process can be iterated until $\hat{V}$ converges.

We remark that setting $K_i = I_{n_i} - X_i^T(X_i^T X_i)^{-1} X_i$ and $Y_i^* = K_i Y_i$ one can derive the likelihood for a single y_i namely

$$\ell_R(V_i; y_i) = -\frac{1}{2}\left[\log|V_i| + \log|X_i V_i X_i| + (y_i - X_i\hat{\beta}_i)^T V_i^{-1}(y_i - X_i\hat{\beta}_i)\right].$$

This can be optimized to get $\hat{V}_i$ which can be used to define a new K_i so as to generate a new $\hat{\beta}_i$ and hence a new likelihood for y_i from which to find a new $\hat{V}_i$, cycling until convergence.

To use (4.4) to get fitted values, we must also estimate D. This is complicated but can be done, see [41], Chap. 5.3 and 9.2. The basic idea is to use the likelihood function for (β, D) given the Y_i's. That is, write

$$L(\beta, D|y_1, \ldots, y_n) = \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} p(y_{ij}|u_i, \beta) p(u_i|D) du_i$$

for the likelihood given by the marginal distribution of $Y_1, \ldots, Y_n$ obtained by integrating out the U_1^n from (Y_i, U_i) for $i = 1, \ldots n$. The derivatives of $L(\beta, D|y_1, \ldots, y_n)$ with respect to the entries in D can be obtained and by using a Newton-Raphson procedure one can generate an estimate $\hat{D}$ for D; describing this is beyond our present scope but the most standard computational procdure is explained in the documentation on PROC MIXED in SAS, see `http://support.sas.com/documentation/`. The EM algorithm can also be used but is less popular.

When the estimates $\hat{\beta}$, $\hat{V}_i$ and $\hat{D}$ are used in the BLUP $E(U_i|Y_i)$ from (4.4) we get an empirical BLUP

$$\hat{U}_i = \hat{D} Z_i^T \hat{V}_i^{-1} (Y_i - X_i \hat{\beta}) \tag{4.11}$$

though the adjective 'empirical' is often omitted for brevity. Moreover, different estimators may be used to form empirical BLUP's and these may have different properties. For instance, even though ML and REML estimators are usually similar, when they differ substantially, REML should be less biased. It is also important to remember that U_i is a random variable so $\hat{U}_i$ is a predictor of U– even though the observation U_i is not seen by us. This is prediction in the same sense that a residual can be said to predict the outcome of an error term that we do not see. In both cases, we must verify that the predictions, as a collection, are representative of the known properties of the distribution they are thought to come from and associate the individual values from (4.11) to the observations that generated them. Now, we obtain fitted values for Y_i

$$\hat{Y}_i = X_i \hat{\beta} + Z_i \hat{U}_i \tag{4.12}$$

that are sometimes called empirical BLUP's for the Y_i's. However, the $\hat{Y}_i$'s are not predictors: Y_i went into the 'prediction' of $\hat{U}_i$ on which $\hat{Y}_i$ depends and all the values of Y_i went into estimating the β.

Some growth curve models can be regarded as a variant on the fixed effect part of the model described above, i.e., as a variant on fixed effects linear regression, see Sec. 2.2.1. For short time series or serial correlation, [74] suggests

$$Y = XBZ + \epsilon \tag{4.13}$$

as a useful growth curve model. In (4.13), Y is a $p \times n$ matrix where n is the number of subects and p is the number of measurements on each subject. The

design matrix X is $p \times d$ of rank d where d is the number of explanatory variables here taken as $1, t, t^2, \ldots, t^{d-1}$ i.e., the degree of the polynomial in time t so that the regression function function for a subject is

$$y = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_{d-1} t^{d-1}.$$

The matrix Z is also a design matrix, of dimension $r \times n$ and rank r, and is used to separate the parameters for the r treatment groups. The matrix B is $d \times r$ and contains the parameters. The parameters form sets, one for each of the r treatment groups. The error matrix ϵ is $p \times n$. As with Y, the columns of ϵ correspond to subjects. Usually, the columns of ϵ are assumed to be IID $N_p(0, \Sigma)$, where Σ is unknown. The structure of (4.13) therefore permits the pooling of data over groups to provide better estimation of B primarily by better estimation of Σ since it is common to all subjects. In this sense, (4.13) can be regarded as a generalization of ANOVA.

To see what this structure means, consider the case $r = 1$, i.e., one treatment group. Then $Z = (1, \ldots, 1)$ of length n and B is a single column of length d. Setting, $d = 3$ and ignoring ϵ, we get for the first subject that

$$\begin{aligned} y_{11} &= \beta_0 + \beta_1 t_1 + \beta_2 t_1^2 \\ y_{12} &= \beta_0 + \beta_1 t_2 + \beta_2 t_2^2 \end{aligned}$$

and hence

$$y_1 = \begin{pmatrix} y_{11} \\ y_{12} \end{pmatrix} = \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

This means that in general the typical (j, k) entry in X is t_j^{k-1}. Analogous interpretations are possible for $r = 2, 3, \ldots$.

Parameter estimation in this class of models proceeds usually by a generalized least squares approach or by an MLE. For the first one finds

$$\hat{B} = \arg\min \mathsf{trace}\,[(Y - XBZ)(Y - XBZ)] = (X^T X)^{-1} X^T Y Z^T (Z Z^T)^{-1}$$

and

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{B}Z)(Y - X\hat{B}Z)^T.$$

It can be shown that $\hat{B}$ is the best linear unbiased estimate of B. Prediction for new subjects or new times proceeds as in fixed effects linear regression.

For the MLE approach to estimating B, replace trace by a determinant. The optimization can be done and provides estimates for B and Σ, see [74] for details. More general growth curve models are presented in [81]; [47] uses a model with $AR(3)$ error to verify that correcting for autocorrelation improves forecasts.

4.1.2. Predicting new outcomes with linear mixed models

At last we turn to prediction of new outcomes with a linear mixed model as opposed to trying to recover an outcome of U. There are at least five predictive settings that are natural. The first, and simplest, is predicting the next values for an individual for whom we have all past observations. Perhaps the earliest technique for this is described in Sec. 4.4 in [81] and given in more generality in Sec. 4 of [22]. It is based on the standard formulae for extracting conditional distributions from a normal. Indeed, [81] derives the simple conditional expectation for a future outcome of a normal regression model with only a random effect term given past outcomes from the same model. This involves the usual partitioning of a variance matrix as used below. Next, we briefly discuss two other prediction cases that differ in terms of what data is available from which to make a prediction. One is the case that we have first stage data on subjects and want to get predictions for the second stage. The other is that we have first and second stage data on $n-1$ subjects and first stage data on the n-th subject so we want to predict the second stage for the n-th subject. The fourth and fifth prediction techniques are more complicated and involve basis element selection and the Box-Cox transformation respectively.

For the first of the five, we extract a conditional expectation as a predictor. Recall (4.2) and assume the error term follows an $AR(1)$ process, a special case of which is IID (when the AR parameter is zero; see [81] Sec. 4.5). Partition Y_i into two parts: $Y_i = (Y_{i1}^T, Y_{i2}^T)^T$ in which Y_{i1} corresponds to the first stage measurements we already have and Y_{i2} corresponds to the second stage measurements we want to predict. Then, X_i can be partitioned into $X_i = (X_{i1}^T, X_{i2}^T)^T$, Z_i can be partitioned into $Z_i = (Z_{i1}^T, Z_{i2}^T)^T$ and $\epsilon_i = (\epsilon_{i1}^T, \epsilon_{i2}^T)^T$ to give

$$\begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} = \begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} \beta + \begin{pmatrix} Z_{i,1} \\ Z_{i,2} \end{pmatrix} U_i + \begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \end{pmatrix}, \tag{4.14}$$

in which β and U remain unchanged. So, $EY_i = ((X_{i,1}\beta)^T, (X_{i,2}\beta)^T)^T$. This induces a block structure on the covariance matrices. Recall, the covariance matrix of Y_i is $V_i = \mathrm{Var}(Y_i) = Z_i D Z_i^T + \sigma^2 I_{n_i}$ which can be partitioned as

$$V_i = \begin{pmatrix} V_{i,11} & V_{i,12} \\ V_{i,21} & V_{i,22} \end{pmatrix} = \begin{pmatrix} Z_{i,1} D Z_{i,1}^T + \sigma^2 I_{n_i,1} & Z_{i,1} D Z_{i,2}^T \\ Z_{i,2} D Z_{i,1}^T & Z_{i,2} D Z_{i,2}^T + \sigma^2 I_{n_i,2} \end{pmatrix}, \tag{4.15}$$

where $I_{n_i,v}$ for $v = 1, 2$ is the identity matrix of dimension equal to that of the first or second part of Y_i. Since Y_i is normal, it is fully specified by the mean and covariance structure we have identified.

So, if we have one observation, say $y_{i,1}$, based on $X_{i,1}$ and $Z_{i,1}$, and we know X_{i2} and Z_{i2}, then the BLUP $\hat{Y}_{i2}$ for Y_{i2} using the outcome y_{1i} of Y_{1i} is

$$\hat{Y}_{i2} = E(Y_{i2}|Y_{i1} = y_{i1}) = X_{i2}\hat{\beta} + Z_{i2}\hat{U}_{i,1}, \tag{4.16}$$

where $\hat{\beta}$ is an estimate of β (using $\hat{V}_i$ found by REML) and $\hat{U}_{i,1}$ is a predictor of U_i formed from the model $Y_{i,1} = X_{i,1}\beta + Z_{i,1}U_i + \epsilon_{i,1}$. (However, the Newton-Raphson and EM algorithm methods for estimating D, needed to find $\hat{U}_{i,1}$, have

been omitted here.) A more general form of (4.16), namely

$$\hat{Y}_{i2} = E(Y_{i2}|Y_{i1} = y_{i1}) = X_{i2}\hat{\beta} + Z_{i2}\hat{U}_{i,1} + \hat{\epsilon}_{i2}. \tag{4.17}$$

holds when the ϵ is $AR(1)$; see [22]. In (4.17), we still have $\hat{U} = E(U_i|Y_{i1} = y_{i1})$ but its interpretation changes becasue of the $AR(1)$ error. We also have $\hat{\epsilon}_{i2} = E(\epsilon_{i2}|Y_{i1} = y_{i1})$ which is nonzero. In fact, [22] gives that

$$\hat{\epsilon}_{i2} = R_{21}R_{11}^{-1}(y_{i1} - X_{i1}\beta - Z_{i1}\hat{U}),$$

and identifies the form of $R_{21}R_{11}^{-1}$ in terms of the $AR(1)$ parameter.

As a second predictive setting suppose all n_i's are the same and we have first stage data on all n subjects so the goal is to predict the second stage of all n subjects. This means we can pool all the first stage data to get improved predictors. Thus, we pool the data to get a better estimator $\hat{\beta}$ for use in (4.16) as well as getting improved estimators of V_i, and D to get a better predictor $\hat{U}_i$ of U_i. Now, (4.16) can be used n times, once for each subject, to get n predictions for the second stages of the n subjects. If we only want to predict the second stage outcome for one of the subjects, the task is of course easier.

A third predictive setting is to imagine that we have complete data, i.e., first and second stage, on $n-1$ subjects and first stage data on the n-th subject. So, our goal is to use all the data to predict the second stage outcome of the n-th subject. One way to proceed is to ignore the second stage data for the first $n-1$ subjects so the prediction problem reverts to the second setting. A better way to proceed is to write (4.14) and (4.15) for the first $n-1$ subjects, i.e., for $i = 1, \ldots, n-1$, and then use one more copy of (4.14) and (4.15) for the n-th subject to form one vector $Y = (Y_{1,1}^T, Y_{1,2}^T, \ldots, Y_{n,1}^T, Y_{n,2}^T)^T$. Then, using the conditioning properties of the normal derive a version of (4.16) for the n-th subject by regarding $Y_{n,2}$ as the last component of the normally distributed vector Y. That is, condition on $Y_{n,1}$ and integrate out the $(Y_{i,1}, Y_{i,2})$s for $i = 1, \ldots, n-1$ to find the conditional expectation $E(Y_{n,2}|Y_{n,1} = y_{n,1})$. It remains to estimate β, R, D (and hence V), and to find $\hat{U}_n$ using all the data available. This can be done by extensions of the techniques here. Further discussion is beyond our present scope, but see [80] for a related problem.

A fourth and more sophisticated setting for forecasting originates in [87]; variants have been explored in [57] and [83]. The core idea of the Shi-Weiss-Taylor (SWT) approach is the following. For each subject i, imagine a curve $Y_i(t) = f(t) + S_i(t) + \epsilon_{it}$ in which $f(t)$ is the population mean profile, $S_i(t)$ is the difference between the population mean profile and the subject-specific profile for subject i, and ϵ_{it} is measurement error. Let us represent this as

$$Y_i = X_i\beta + Z_iU_i + \epsilon_i^*, \tag{4.18}$$

where Y_i is an $n_i \times 1$ vector of measurements at $t_i = (t_{i1}, \ldots, t_{in_i})$, X_i and Z_i are, respectively, $n_i \times J$ and $n_i \times K$ matrices with in which their j-th columns are the j-th B-spline basis element evaluated at the values $t_{i1}, \ldots, t_{in_i}$. So, the only difference between X_i and Z_i is that the first uses J B-spline basis elements

and the second uses K B-spline basis elements. Clearly, $X_i\beta$ and Z_iU_i are approximations to f and S_i, respectively, for the elements in t_i. In addition, we assume $U_i \sim N(0, D_K)$ where D_K is $K \times K$ and ϵ_i^* is $N_{n_i}(0, \sigma^2 I_{n_i})$, a new error term hopefully behaving similarly to the original ϵ_{it}. Thus, making predictions from (4.18) requires us to choose J and K as well as estimate β, D, and σ and obtain the $\hat{U}_i$'s.

To simplify the problem, SWT use a principal components decomposition for D_K to reduce the number of parameters that must be estimated. Write $D_K = \Delta\Lambda\Delta^T$ where Δ is the orthogonal matrix of normalized eigenvectors Δ_k of D_K and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_K)$ with eigenvalues in decreasing order. Since we are going to choose a serviceable K, it is helpful to set $\delta_{ik} = \lambda_k^{-1/2}\Delta_k^T U_i$ and $C_{ik} = \lambda_k^{1/2} Z_i \Delta_k$ $(n_i \times 1)$ and represent Z_iU_i as

$$Z_iU_i = \sum_{k=1}^{K} (Z_i\lambda_k^{1/2}\Delta_k)(\lambda^{-1/2}\Delta_k^T U_i) = \sum_{k=1}^{K} C_{ik}\delta_{ik}.$$

It is seen that $\delta_{ik} \sim N(0,1)$ and the directions Δ_k corresponding to higher eigenvalues are more important to the approximation of S_i.

Looking component-wise at the vector Y_i, it is reasonable to find K (for large enough J) so that for any $i = 1, \ldots n$ the real valued function

$$Y_i(t) \approx B(t)\beta + \sum_{k=1}^{K} \delta_{ik} C_k(t) + \epsilon_{it}^* = B(t)\beta + s_{iK}(t) + \epsilon_{it}^* \tag{4.19}$$

representing a generic entry of Y_i for a generic t_{ij} (written as t), provides the best fit to the data. In (4.19), we have assumed all n_is are the same so that $B(t)$ $(J \times 1)$ is the vector of evaluations of the first J spline functions at $t \in \mathbb{R}$, i.e., $B(t)$ is a generic row of any X_i independent of i. Likewise, $C_k(t)$ (1×1) is a component of $\lambda_k^{1/2} Z_i \Delta_k$ formed by using a generic row of Z_i and in which the spline basis elements are evaluated at a general $t \in \mathbb{R}$, i.e., $C_k(t)$ is a generic form of C_{ik} independent of i.

Clearly, the random effects portion is $s_{iK}(t)$ and the appropriateness of a given K can be evaluated by examining the variance explained. Note that the same basis elements are used in both the fixed and random effects portions of the model (though this need not be the case) and there is no harm in regarding the J variables in the fixed portion separately from the K variables in the random effects portion. It is seen that both $B(\cdot)$ and the $C_k(\cdot)$s depend on the number and location of knots, with each C_k contributing one random effect.

To implement this model, one can choose the number of knots via variance-bias tradeoff. More knots give a smaller bias but greater variance; fewer knots give a higher bias but a smaller variance. The location of the knots matters as well: [87] suggests using some knots where data were collected and others where there is more curvature in the response.

There are several ways to choose J and K, and [87] give four: Using a likelihood ratio test, using cross-validation, examining the role of the $C_{ik}(t)$ in the

reduction of estimates of $\hat{\sigma}$, and, for a fixed J, choosing K by looking at the relative percentage of the variation explained by the random effects (this amounts to looking at a form of R^2).

Since $R = \sigma^2 I_{n_i}$, the parameters and the outcomes of the random effects terms, i.e., $\hat{\beta}$, $\hat{D}_K$, $\hat{U}_i$, and $\hat{\sigma}$, can be estimated by a variety of methods including MLE's, REML's and Bayesian methods. The SWT method starts by fixing J and fitting (4.18) using $Z_i = (B(t_{i1}), \ldots, B(t_{in_i}))^T$. Now, D_K is the $K \times K$ covariance matrix of U_i and can be estimated by the MLE or by the EM algorithm. The eigenvalues and eigenvectors of D_K are combined with the B-spline basis to obtain the transformed B-splines $C_k(t)$'s. The second step is to do model selection (over K) using the C_k's. So, refit (4.18) for $K = 1, 2, \ldots$ and use cross-validation scores, the log-likelihood, $\hat{\sigma}$ and R^2. This can be done using an EM algorithm so that the four methods can be used to choose K.

For given K and i, the regression function can be written as

$$\hat{y}(t) = \hat{f}(t) + \hat{s}_i(t) = B(t)\hat{\beta} + \sum_{j=1}^{K} \hat{u}_{ij} C_j(t),$$

where the $\hat{u}_{ij}$ are predictions for the U_{ij} (which were transformed to δ_{ij}s) with residuals $\hat{e}_i(t) = y_i(t) - \hat{f}(t) - \hat{s}_i(t)$. Then the natural estimate of the population curve is given by confidence bands of the form

$$\hat{f}(t) - z_{\alpha/2}\sigma_{f(t)} \leq f(t) \leq \hat{f}(t) + z_{\alpha/2}\sigma_{f(t)},$$

where

$$\sigma_f(t) = B^T(t) \left(\sum_{i=1}^{n} X_i^T (\hat{\sigma}^2 I_{n_i} + Z_i \hat{D}_K Z_i^T) X_i \right)^{-1} B(t).$$

In addition, [87] identifies the quantile curves associated with (4.19). If all the n_i's and t_i's are the same then $C_{ik} = C_k$ and they too can be regarded as functions of t since the J entries are based on spline functions. Writing $C(t) = (C_1(t), \ldots, C_K(t))$, the $100\nu\%$ quantile curve is estimated by

$$\hat{Y}_\alpha(t) = B(t)\hat{\beta} + z_\alpha \left(C^T(t) \hat{D}_K C(t) + \hat{\sigma}^2 \right)^{1/2}, \tag{4.20}$$

for any t. This gives a $(1-\alpha)100\%$prediction interval for a future value of $Y(t)$ of the form $[\hat{Y}_{\alpha/2}, \hat{Y}_{1-\alpha/2}]$. By the symmetry of the normal, the median curve with $\alpha = .5$ will be the same as the conditional expectation, see (4.12) and (4.16), which are BLUP's.

For contrast with SWT, we next describe a fifth predictive technique for longitudinal settings due to [27]. The [27] method can be regarded as a simpler version of SWT and, while a bit ad hoc, it can give decent results in examples (apart from boundary effects and some problems with non-uniform smoothing as might be conjectured once the procedure is explained). At root, the approach

in [27] is a growth curve model but of a different form from that in [74] described at the end of Sec 4.1.1. The central idea is to find curves like (4.20) that give time dependent prediction intervals. Again, the endpoints are percentiles from the normal distributions describing the response at each time or one could choose the median curve as a point-predictor.

For a single variable of interest Y recall the Box-Cox transformation and transform Y to X:

$$x = \frac{(y/\mu)^\lambda - 1}{\lambda} \quad \lambda \neq 0$$

and $x = \ln(y/\mu)$ when $\lambda = 0$; a useful discussion of the Box-Cox transformation, including the effect of influential data, can be found in [86]. The median of Y maps to $x = 0$; the standard deviation (SD) of X, say σ, is the coefficient of variation of Y and λ is chosen to minimize the SD of X. This gives

$$z = \frac{x}{\sigma} = \frac{(y/\mu)^\lambda - 1}{\lambda\sigma} \quad \lambda \neq 0$$

and $z = (\ln(y/\mu))/\sigma$ for $\lambda = 0$. So, it is hopefully safe to assume $Z \sim N(0, 1)$.

Let us extend the Box-Cox transformation to include an explanatory variable, say t. If Y depends on t then the optimal λ may depend on t and therefore so will μ and σ. Let $L(t)$, $M(t)$ and $S(t)$ represent the curves for λ, μ and σ as t varies. Now, the Box-Cox transformation takes the form

$$z(t) = \frac{(y(t)/M(t))^{L(t)} - 1}{L(t)S(t)} \quad , \quad L(t) \neq 0, \tag{4.21}$$

and $z(t) = (\ln(y(t)/M(t)))/S(t)$, when $L(t) = 0$. So, for $L(t) \neq 0$, (4.21) gives

$$\hat{Y}_\alpha(t) = M(t)(1 + L(t)S(t)z_\alpha)^{1/L(t)}, \tag{4.22}$$

and when $L(t) = 0$,

$$\hat{Y}_\alpha = M(t)e^{S(t)z_\alpha} \tag{4.23}$$

from which quantile based prediction intervals can be obtained; setting $\alpha = .5$ gives a median point predictor.

The remaining difficulty in implementing this is estimating $L(t)$, $M(t)$ and $S(t)$. In (4.21), we can assume Z is $N(0, 1)$ and following [27] obtain the likelihood function

$$\ell = \ell(L, M, S) = \sum_{i=1} \left(L(t_i) \ln \frac{y_i}{M(t_i)} - \ln S(t_i) - \frac{1}{2} z_i^2 \right)$$

(neglecting the constant) where $z_i = x_i/S(t_i)$ and $x_i = ((y_i/M(t_i))^{L(t_i)} - 1)/L(t_i)$ for independent observations $i = 1, \ldots, n$ (analogous to 4.21). Now, consider the penalized likelihood

$$\ell^* = \ell - \alpha_\lambda \int (L''(t))^2 dt - \alpha_\mu \int (M''(t))^2 dt - \alpha_\sigma \int (S''(t))^2 dt, \tag{4.24}$$

where α_λ, α_μ and α_σ are smoothing parameters. Maximizing (4.24) leads to cubic splines with knots at the distinct t_i's; see [26] Chap. 3.2, so that only the three smoothing parameters remain to be chosen. [27] uses an iterative procedure to get results by choosing initial values of α_λ, α_μ and α_σ, maximizing in (4.24), and then empirically choosing new values for α_λ, α_μ and α_σ. The net result is three smoothing spline estimates $\hat{L}(t)$, $\hat{M}(t)$ and $\hat{S}(t)$ which can be used directly in (4.22) or (4.23). More conservatively, confidence bounds on $\hat{L}(t)$, $\hat{M}(t)$ and $\hat{S}(t)$ can be obtained and used in (4.22) or (4.23).

4.2. Generalized linear models and estimating equations

A generalized linear model (GLM) – not to be confused with the general linear model – is actually a family of models, the element of the family being determined by the nature of the data. Logistic regression for binary variables, see Sec. 2.2.3, is an example of a generalized linear model but there are many others. We start our description with the case that the Y_is are independent even though this is unrealistic for longitudinal data; this will be relaxed shortly. To begin, there are three defining features for a GLM:

- A linear predictor $\eta_i = x_i^T\beta$ where x_i is the vector of explanatory values for subject i and β is the parameter vector.
- A link function g specifying the relationship between $\mu_i = E(Y_i)$, the expected value of subject i at x_i, and the linear predictor. That is,

$$g(\mu_i) = g(E(Y_i)) = \eta_i = x_i^T\beta.$$

- A relationship between the conditional variance of Y_i and the covariates,

$$\mathrm{Var}(Y_i) = \phi_i V(\mu_i)$$

  where ϕ_i is a possibly subject dependent scaling parameter that is either known or to be estimated and $V(\mu_i)$ is a known variance function.

For exponential families, the link function is chosen to be the transformation of the mean so as to get a linear predictor, i.e., the link is determined by the parametric family. So, for the case of logistic regression, η_i is used with the logit function $g(u) = \log(u/(1-u))$ (see (2.23)) and $V(\mu_i) = \mu_i(1-\mu_i)$, effectively setting all $\phi_i = 1$.

A different example arises if we consider count data. Recall that for $Y \sim Poisson(\lambda)$, $\mu = E(Y) = \mathrm{Var}(Y)$. If Y_i is the count for subject i with explanatory variables x_i then we use $\eta_i = x_i^T\beta$ with

$$\log E(Y_i) = \log \mu_i = x_i^T\beta.$$

To model the variance we set

$$\mathrm{Var}(Y_i) = \phi_i E(Y_i),$$

where $\phi_i > 1$ and $\phi_i < 1$ represent over- and under-dispersion (relative to the Poisson) respectively; often a common value ϕ for the ϕ_i's is assumed. Technically, if $\phi \neq 1$, the underlying distribution of Y is not Poisson, but this procedure is usually called Poisson regression anyway. Poisson regression can be generalized to log-linear models but we do not describe these here. The probit link, based on the inverse of the standard normal distribution function, is also used in settings where Y is binary and the explanatory variables can be taken as normal. It is sometimes regarded as quite similar to the logit link. There are numerous other link functions, many of which have been implemented in **R**.

Provided the GLM model has been specified, one can estimate β by solving the estimating equation

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T \mathrm{Var}(Y_i)^{-1}(y_i - \mu_i) = 0. \tag{4.25}$$

This equation only uses the first two moments of Y_i and so applies to any distribution with those moments, i.e., we do not have to assume a specific distributional form for the Y's. The details of how to use estimating equations, and their properties, are beyond our present scope. It is enough to observe that the estimates of β are typically consistent, asymptotically normal, although not necessarily efficient. Techniques for estimating the ϕ_i's i.e., when they cannot all be taken as one, are beyond our present scope, however in many cases they can be found computationally.

Turning to prediction, in principle, we can examine each instance of a GLM model and carefully derive a correct PI given g (but see [75] for a computational approach). Since this does not seem to have been expressed mathematically, we proceed (with great informality) to describe a procedure that may work for any g. It rests on asymptotics; details for specific g's are beyond our present scope.

First we look at point predictors. Given that the parameters have been estimated, the natural point predictor for a new value x_{n+1} is $\hat{Y}_{n+1} = g^{-1}(x_{n+1}\hat{\beta})$. This was seen for logistic regression in Sec. 2.2.3. Obviously, $\hat{Y}_{n+1}$ is also the natural point estimator for μ_{n+1} however its roles in prediction and estimation are different. So, for greater accuracy we take into account the possible nonlinearities in g and find a modification of $\hat{Y}_{n+1}$.

If we naively use a second order Taylor expansion we can write

$$\begin{aligned} Y_{n+1} &\approx g^{-1}(\eta_{n+1}) + (g^{-1}(\hat{\beta}x_{n+1}) - g^{-1}(\eta_{n+1})) \\ &\approx g^{-1}(\eta_{n+1}) + (\hat{\beta}x_{n+1} - \eta_{n+1})(g^{-1})'(\eta_{n+1}) \\ &+ (\hat{\beta}x_{n+1} - \eta_{n+1})^2(g^{-1})''(\eta_{n+1})/2. \end{aligned} \tag{4.26}$$

Taking expectations, the middle term on the right is zero giving

$$\begin{aligned} E(Y_{n+1}) &\approx g^{-1}(\eta_{n+1}) + E(\hat{\beta}x_{n+1} - \eta_{n+1})^2(g^{-1})''(\eta_{n+1})/2 \\ &\approx g^{-1}(\eta_{n+1}) + (1/2)(g^{-1})''(\eta_{n+1})x_{n+1}^T \mathrm{Var}(\hat{\beta})x_{n+1}. \end{aligned}$$

So, plugging in $\hat{\eta}_{n+1} = x_{n+1}\hat{\beta}$ for η_{n+1} and finding an estimate for $\mathrm{Var}(\hat{\beta})$ will give a useful point predictor – or point estimator – $\hat{Y}_{n+1}$ for Y_{n+1}, provided the mean is a reasonable summary for the location of the distribution.

To derive an interval predictor we begin by noting that in the formulation of the predictor we will want to use $\hat{Y}_{n+1}$ as an estimator for $E(Y_{n+1})$. Moreover, the natural way to find an standard error is via the delta method. To implement this, ignore the second derivative term in (4.26) and take variances on both sides. Since the cross-term drops out, this gives

$$\begin{aligned} \mathrm{Var}(Y_{n+1}) &= \mathrm{Var}(g^{-1}(\eta_{n+1})) + \mathrm{Var}((\hat{\beta}x_{n+1} - \eta_{n+1})(g^{-1})'(\eta_{n+1})) \\ &= \mathrm{Var}(\hat{\beta}x_{n+1}(g^{-1})'(\eta_{n+1})) \\ &= ((g^{-1})'(\eta_{n+1}))^2 x_{n+1}^T \mathrm{Var}(\hat{\beta}) x_{n+1}. \end{aligned} \tag{4.27}$$

Now, one can invoke asymptotic normality and use

$$\hat{Y}_{n+1} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}(Y_{n+1})}} \tag{4.28}$$

as a confidence interval for $E(Y_{n+1})$ since β and $\mathrm{Var}(\hat{\beta})$ can be estimated (and x_{n+1} is known).

Now we can specify interval predictors. Let us write $Y_{n+1} \sim G_{\eta_{n+1}}$ where $G_{\eta_{n+1}}$ is the distribution of Y_{n+1} given $\eta_{n+1} = g^{-1}(x_{n+1}\beta)$. For a GLM, $G_{\eta_{n+1}}$ is taken to be a known exponential family. So, in principle, we can obtain expressions for $\rho_{\alpha/2,\eta_{n+1}}$ and $\rho_{1-\alpha/2,\eta_{n+1}}$ the $\alpha/2$ and $1-\alpha/2$ percentiles of $G_{\eta_{n+1}}$, respectively. It remains to estimate the percentiles; this devolves to estimating β. We could therefore just plug in an estimate $\hat{\eta} = x_{n+1}\hat{\beta}$ for $\eta_{n+1} = x_{n+1}\beta$. Or, to be conservative, we could use a lower confidence bound on η_{n+1} in $\rho_{\alpha/2,\eta_{n+1}}$ and an upper confidence bound on η_{n+1} in $\rho_{1-\alpha/2,\eta_{n+1}}$. These can be naturally taken to be of the form $\hat{\beta}x_{n+1}z_{\alpha/2}x_{n+1}^T\mathrm{Var}(\hat{\beta})x_{n+1}$ and $\hat{\beta}x_{n+1} + z_{1-\alpha/2}x_{n+1}^T\mathrm{Var}(\hat{\beta})x_{n+1}$. Furthermore, writing $g^{-1}(\eta_{n+1}) = E(Y_{n+1})$ lets us transform confidence intervals for $E(Y_{n+1})$ via (4.28) to confidence intervals on η_{n+1} for use in $\rho_{\alpha/2,\eta_{n+1}}$ and $\rho_{1-\alpha/2,\eta_{n+1}}$.

To use GLM models for longitudinal data one extends them to Generalized Estimating Equation models (GEE's). For $i = 1,\ldots,n$ subjects measured at $j = 1,\ldots,m$ timepoints we write the defining conditons above as $\eta_{ij} = x_{ij}^T\beta$, $E(Y_{ij}) = \mu_{ij}$, $g(\mu_{ij}) = \eta_{ij}$ and $\mathrm{Var}(Y_{ij}) = \phi V(\mu_{ij})$. Next, one needs a 'working correlation' $R_i(\theta)$ where θ indicates the parameters defining the entries in the $m \times m$ matrix R_i for the i-th subject. Clearly, $R_i = I_m$, the $m \times m$ identity matrix is unreasonable – it would mean the observations on a given subject were uncorrelated defeating the point of a longitudinal analysis. Sometimes $R_i(\theta)$ is taken to be the matrix with all entries ρ, meaning any two measurements on a subject have the same correlation; other choices are possible.

To use the correlation matrices, the R_i's, we must convert them to a covariance structure. By definition of a GLM, the form $V(\cdot)$ is known so let $A_i = \mathrm{diag}(V(\mu_{i1}),\ldots,V(\mu_{im}))$, the $m \times m$ diagonal matrix with entries given

by the variances of the Y_{ij}'s for $j = 1, \ldots, m$. Then, set

$$V_i(\theta) = \phi A_i^{1/2} R_i(\theta) A_i^{1/2}.$$

Now, the analog of (4.25) is

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T V_i(\hat{\theta})^{-1} (y_i - \mu_i) = 0, \tag{4.29}$$

in which $y_i = (y_{i1}, \ldots, y_{im})^T$, $\mu_i = (\mu_{i1}, \ldots, \mu_{im})^T$, and $\partial \mu_i / \partial \beta$ is the $m \times p$ matrix with (j,k) entry $\partial \mu_{ij} / \partial \mu_k$ for $j = 1, \ldots, m$ and $k = 1, \ldots, p$. Expression (4.29) can be used to estimate β provided we have consistent estimates $\hat{\theta}$ and $\hat{\phi}$ of θ and ϕ. The resulting estimates are often called quasi-likelihood estimators because they only use the first two moments of Y_{ij} leaving the rest of the distribution unspecified. Roughly, one uses $\hat{\theta}$ and $\hat{\phi}$ to get $\hat{\beta}$ which is then used to get new estimates of θ and ϕ, cycling until convergence. A variance for $\hat{\beta}$ can also be give. Often, it is enough to use

$$\hat{V}(\hat{\beta}) = \sum_{j=}^{m} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T V_i(\hat{\theta}) \left(\frac{\partial \mu_i}{\partial \beta} \right),$$

but usually better estimtes can be given.

Even given successful parameter estimation, prediction in the context of GEE's is problematic. The reason is that the GEE models only use assumptions about the first and second moments. This is often satisfactory for parameter inference but not obviously for point predictions or prediction intervals which may depend on higher order moments. Indeed, if one naively uses a GEE model as if the corresponding GLM model (with the appropriate covariance structure) were true, then it is not clear how wide a class of fully specified models are being represented by the first two moments of the GLM model. More forcefully, it is not clear how far the true model may be from the GLM model while still having the same first two moments. It may be that two different parametric families that lead to the same GEE model require quite different predictors. As a generality, it seems this issue remains to be studied.

We conclude this subsection with the observation that there seems to be little work on prediction with generalized linear models even in the case that they are treated as an extension to the general linear model i.e., without taking a dependence structure into account as needed for longitudinal data. There are a few notable exceptions such as [91] and [71], but, overall, prediction in GLM's – and especially in GEE's – seems relatively unexplored.

4.3. Generalized linear mixed models

Just as generalized linear models include a link function to express the conditional mean of a response given covariates in terms of a fixed effects linear

model, generalized linear mixed models (GLMM's) introduce a link function to express the conditional mean of the response given the covariates and random effects in terms of a mixed effects linear model.

The basic idea is that some function of the conditional mean of Y_{ij} given the random effects U_i and explanatory variables X_i is representable as a linear mixed model and that conditional on the U_i, the Y_{ij} are independent. So, for Y_{ij}, the j-th measurement on subject i, we assume there is a vector U_i of length q so that $(Y_{ij}|U_i)$ belongs to an exponential family. A necessary assumption is that $\mathrm{Var}(Y_{ij}|U_i)$ is a function of $E(Y_{ij}|U_i)$ so that estimating the variance is feasible. Now, for some function g, we must assume that

$$g(E(Y_{ij}|U_i)) = \sum_{k=1}^{p} X_i(j;k)\beta_k + \sum_{k=1}^{q} Z_i(j;k)U_{ik} \tag{4.30}$$

and that the U_i are assigned some distribution. In the linear mixed model case, $\mathrm{Var}(Y_{ij}|U_i) = \sigma^2$, $g \equiv 1$, and $U_i \sim N(0, D)$.

The logistic regression in Sec. 2.2.3 can be regarded as a GLM and extended to a GLMM as follows. Suppose Y_{ij} is binary taking values 0 and 1. Then, conditional on the U_i's, the Y_{ij}'s are Bernoulli$(E(Y_{ij}|U_i))$ so $\mathrm{Var}(Y_{ij}|U_i) = E(Y_{ij}|U_i)(1-E(Y_{ij}|U_i))$. The natural model is (4.30). So, if $q = 1$, and $Z_i(j,1) = 1$ we get

$$g(E(Y_{ij}|U_i)) = \sum_{k=1}^{p} X_i(j;k)\beta_k + U_{i1}$$

and assuming g is a logit function

$$\log \frac{P(Y_{ij}=1|U_i)}{P(Y_{ij}=0|U_i)} = \sum_{k=1}^{p} X_i(j;k)\beta_k + U_{i1}.$$

Another example is Poisson regression. Suppose $Y_{ij} \sim \mathrm{Poisson}(\lambda_{ij})$ and that the $E(Y_{ij}|U_i)$ are independent for $j = 1, \ldots, n_i$, where U_i is the random component for subject i. Then, $E(Y_{ij}|\lambda_{ij}) = \lambda_{ij}$. So, we can write

$$\log(\lambda_{ij}) = X_i(j)\beta + Z_i(j)U_i$$

and assign a distribution to the U's such as $U \sim N(0, D)$.

Estimation in GLMM's is, to some extent, still a research topic although some techniques are gaining acceptance. Likelihood methods have been developed, see [10] for one example, as have least squares techniques, see [102] for an instance.

Predicting new observations in GLMM's has also been studied, see for instance [9]. However, there seem to be relatively few references for the general case of predicting new outcomes even for specific choices of link function. In a spatial statistics context, [95] Chap. 9, Sec. 7.7, p. 433, states that using a pseudo-likelihood approach for predicting new observations in a GLMM is similar to kriging with the GLM mean and variance structure and gives several references. For a Bayesian version see [102] and the references therein. However, this material is beyond our present scope.

4.4. Nonlinear mixed models (NLMM)

One important early contribution to NLMM methodology is [68] who proposed a model that can be regarded as a subclass of the class to be presented below. This general class of NLMM's saw rapid methodological development in the late 90's and early 00's; this is well summarized in [32]. Using the hierarchical form of the mixed effects model they define the individual level as

$$Y_{ij} = f(X_{ij}, \beta_i) + \epsilon_{ij}$$

for $j = 1, \ldots, n_i$ in which X_{ij} contains the explanatory variables for subject i and response j, β_i is a collection of parameters specific to individual i, and the form of f is known but may depend on β_i. It is understood that X_{ij} has two parts, t_{ij} and u_i, where t_{ij} can be regarded as time and u_i taken to represent other conditions (e.g., initial dose of a drug). Moreover, at each t_{ij}, the individual deviations satisfy

$$E(\epsilon_{ij}|u_i, \beta_i) = E(Y_{ij} - f(x_{ij}, \beta_i)|u_i, \beta_i) = 0.$$

The population level model is given by a function d satisfying

$$\beta_i = d(a_i, \beta, U_i)$$

depending on p fixed effects β, characteristics a_i of subject i that are independent of anything else, and q random effects in U_i. The population level model describes how β_i varies over individuals as a result of individual attributes a_i, population quantities β, and individual variation in U_i. Usually, $E(U_i|a_i) = E(U_i) = 0$ and $\mathrm{Var}(U_i|a_i) = \mathrm{Var}(U_i) = D$. As before, it is common to take $U_i \sim N(0, D)$. In a slightly more general version of this hierarchical structure for NLMM's, [32] Secs. 3.5 and 3.6 discusses Bayesian analysis and individual inference. However, the details are beyond our present scope.

An important recent contribution is [76]. They provide an extension of the EM algorithm to multilevel NLMM's. The model class they use differs from that of [32] due to an extra layer of variability, i.e., in addition to within- and between-subject variability there may be a grouping of subjects. The details in [32] are beyond our present scope, as is the general topic of multilevel NLMM's. Altogether, prediction in these settings (GLM's, GEE's, GLMM's, and NLMM's) seems relatively unexplored.

5. Survival analysis

In its basic form, the key question asked in survival analysis is 'How long will a given subject last without change?'. If the survival time for a randomly chosen subject is denoted $Y \geq 0$ then we want to know the distribution P of Y so that we can make predictions about future subjects from the same population. As in Sec. 2.1, we might have n outcomes $y_1, \ldots, y_n$ of Y and want to predict Y_{n+1}. If we use the mean $\bar{y}$, or some other point predictor, we can form prediction

intervals as in Sec. 2.1.1 or 2.1.3, whether $\sigma^2 = \mathrm{Var}(Y_i)$ is known or must be estimated. Rather than using Chebyshev's inequality and other standard results to quantify the behavior of a point predictor, one can try to understand the whole distribution of the lifetimes, in particular how it depends on explanatory variables, if they exist. This conveys much more information.

In this section, we are only presenting the classical results; Bayesian treatments of survival analysis abound and one highly readable reference is [56]. Also, we do not look at time dependent variables, missing data, complex censoring, or competing risk models.

5.1. Using the distribution of survival times for prediction

The distribution of survival times is usually treated in complementary form. That is, if Y has distribution function F then its survival function $S(\cdot)$ is $1-F(\cdot)$. That is,

$$S(y) = P(Y \geq y),$$

the probability that a given subject has lifetime as long as or longer than y. In this subsection we discuss estimating S non-parametrically and, briefly, how to convert an estimate of S into a diagnostic model.

5.1.1. The Kaplan-Meier estimator

The standard nonparametric estimator of S is called the Kaplan-Meier (KM) or product-limit estimator, see [61], and is particularly useful because it accommodates right-censored data, reducing to the usual histogram when the data are not censored. The most compact form of this estimator expresses its values as a growing product:

$$\hat{S}(y) = \begin{cases} 1 & \text{if } y < y_{(1)}, \\ \prod_{y_i \leq y}[1 - \frac{d_i}{n_i}] & \text{if } y_{(1)} \leq y. \end{cases} \tag{5.1}$$

where $y_{(i)}$ is the i-th order statistic from $y_1, \ldots, y_n$'s, n outcomes of Y, and the d_i's are the number of events (e.g., deaths) at time $y_{(i)}$. In the absence of censoring, n_i is the number of survivors just before y_i. If some events are right censored then n_i is the number of survivors less the number of losses due to censoring just before y_i. This makes sense because it is only the survivors who are still being observed who are at risk of death. The main $\mathcal{O}(1/\sqrt{n})$ pointwise weak consistency result with an identified covariance is due to [13], see Sec. 5 and 6. (There are important contributions predating [13]; Greenwood's formula used below dates from 1926!) However, uniform weak consistency (with a $\mathcal{O}(1/\sqrt{n})$ rate) is attributed to [51] and strong uniform consistency (with the rate incompletely handled but decreased from $\mathcal{O}(1/\sqrt{n})$ by $(\log\log n)^{1/2}$) is in [46]. Importantly, [21] established uniform weak and strong laws with rates $\mathcal{O}(1/n^q)$ where $q < 1/2$ is determined by the censoring.

Since survival distributions are usually right-skewed, we examine the median of $\hat{S}$ as a predictor for the lifetime of the next subject. This is

$$\widehat{\mathsf{med}}(Y) = \min\{y_i|\ \hat{S}(y_i) \leq .5\}; \tag{5.2}$$

other percentiles can be estimated similarly. Moreover, it is not too hard to give an approximate variance for the estimate of a percentile of $\hat{S}$ so that the results of Sec. 2.1 can be used (see 2.7 and 2.9). Indeed, [28] gives an approximation for the variance of $\widehat{\mathsf{med}}(Y)$ based on using the Taylor series approximation for the variance of a function of a random variable; this is of a different form from – and more useful here than – that given in [82] Sec. 4.1.

Treating $\widehat{\mathsf{med}}(Y)$ as a random variable and $S(\cdot)$ as a function we get

$$\mathrm{Var}(S(\widehat{\mathsf{med}}(Y))) \approx \left(\frac{dS(t)}{dt}|_{\widehat{\mathsf{med}}(Y)}\right)^2 \mathrm{Var}(\widehat{\mathsf{med}}(Y)), \tag{5.3}$$

from which we can solve for the second factor on the right. Indeed, the left hand side of (5.3) is approximated by using Greenwood's formula that for $y \in [y_{(k)}, y_{(k+1)}))$

$$\widehat{\mathrm{Var}}(\hat{S}(y)) \approx \hat{S}(y)^2 \left(\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}\right); \tag{5.4}$$

see [28]. Greenwood's formula is not a good approximation in the tails of the survival function but there are improvements to it and here we are only looking at the median. Next, we can write

$$\frac{dS(t)}{dt}|_{\widehat{\mathsf{med}}(Y)} = -p(\widehat{\mathsf{med}}(Y))$$

where $p(\cdot)$ is the density of P. So, to get an estimate $\widehat{dS(t)/dt}$ of $dS(t)/dt$, it is enough to have an estimate of the density p. That is, we can write

$$\widehat{\frac{dS(t)}{dt}}|_{\widehat{\mathsf{med}}(Y)} = -\hat{p}(\widehat{\mathsf{med}}(Y))$$

where $\hat{p}$ is an estimator of the density of the survival time evaluated at $\widehat{\mathsf{med}}(Y)$. Perhaps the simplest choice is to set

$$-\hat{p}(\widehat{\mathsf{med}}(Y)) = \frac{\hat{S}(\hat{u}) - \hat{S}(\hat{\ell})}{\hat{\ell} - \hat{u}}, \tag{5.5}$$

where $\hat{u} = \max\{y_j|\hat{S}(y_j) \geq 1 - (p/100) + \epsilon\}$, and $\hat{\ell} = \min\{y_j|\hat{S}(y_j) \geq 1 - (p/100) - \epsilon\}$, see [28]. Using (5.4) and (5.5) in (5.3), we get an estimate for $\mathrm{Var}(\widehat{\mathsf{med}}(Y))$. Asymptotic normality can sometimes be invoked to give $1-\alpha$ CI's for the median of the form $\widehat{\mathsf{med}}(Y) \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\mathsf{med}}(Y))}$.

An alternative approach is to construct a CI for the median directly from the lower and upper bounds of the CI for the Kaplan-Meier estimator. An asymptotic $100(1-\alpha)$ CI for the Kaplan-Meier estimator can be constructed using the Greenwood variance estimate as $\hat{S}(y) \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{S}(y))}$, setting $y = \widehat{\mathsf{med}}(Y)$. Note, however, that this interval can extend below 0 or above 1, and have poor coverage properties [90].

One way around getting CI's that contain points outside (0,1) is to use a logarithm or logit transform of survival and transform the resulting CI into a CI for $\hat{S}(y)$, see [28]. For instance, the log-transform would lead us to form $100(1-\alpha)$ CI's for $\log S(y)$ of the form

$$\log \hat{S}(y) \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\log \hat{S}(y))}.$$

Using the Taylor series approximation to the variance of a function of a random variable again leads to

$$\mathrm{Var}(\ln \hat{S}(y)) \approx \frac{\mathrm{Var}(\hat{S}(y))}{\hat{S}(y)^2}$$

in which we can use Greenwood's formula for the numerator. Undoing the log, this leads to a $100(1-\alpha)$ CI for $\hat{S}(y)$ of the form

$$[\hat{S}(y)\exp\left(-z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{S}(y))}/\hat{S}(y)\right), \hat{S}(y)\exp\left(z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{S}(y))}/\hat{S}(y)\right)]. \tag{5.6}$$

The upper and lower limits of this confidence interval can now be used in (5.2) to get upper and lower bounds on $\widehat{\mathsf{med}}(Y)$. (However, the confidence level is not clear because (5.6) is used for several values of y_i.)

Having examined the variance of $\mathsf{med}(Y)$ as an estimator, we turn to using $\widehat{\mathsf{med}}(Y)$ as a predictor for the survival time of the next subject, i.e., to enable use of approaches from Sec. 2.1. The simplest analysis parallels (2.1). Replacing the mean with the median, invoke asymptotic normality and use the fact that the rate of convergence of $\hat{S}$ to S is $\mathcal{O}(1/\sqrt{n})$ uniformly on compact sets. Explicitly, write $\mathrm{Var}(\widehat{\mathsf{med}}(Y)) \approx \sigma_m^2/n$, $\mathrm{Var}(Y_{n+1}) = \sigma^2$, and let $\tau > 0$. Then,

$$\begin{aligned}
& P(|\widehat{\mathsf{med}}(Y) - Y_{n+1}| \geq \frac{\sigma_m/\sqrt{n} + |E\hat{\mathsf{med}}(Y) - EY_{n+1}| + \sigma}{\tau}) \\
\leq\ & \frac{\tau}{\sigma_m/\sqrt{n} + |E\hat{\mathsf{med}}(Y) - EY_{n+1}| + \sigma}\Big(E|\widehat{\mathsf{med}}(Y) - E\widehat{\mathsf{med}}(Y)| \\
+\ & |E\widehat{\mathsf{med}}(Y) - E(Y_{n+1})| + E|E(Y_{n+1}) - Y_{n+1}|\Big) \\
\leq\ & \frac{\tau}{\sigma_m/\sqrt{n} + |E\hat{\mathsf{med}}(Y) - EY_{n+1}| + \sigma}\Big(\sqrt{\mathrm{Var}(\widehat{\mathsf{med}}(Y))} \\
+\ & |E\widehat{\mathsf{med}}(Y) - E(Y_{n+1})| + \sqrt{\mathrm{Var}(Y_{n+1})}\Big) \\
\approx\ & \tau.
\end{aligned} \tag{5.7}$$

So, if σ, σ_m, and $E(\widehat{\mathsf{med}}(Y)) - E(Y_{n+1})$ can be reliably estimated and τ can be chosen small enough that the lower bound on $|\widehat{\mathsf{med}}(Y) - Y_{n+1}|$ is meaningful, we can obtain prediction intervals parallel to (2.2). In a similar way, (2.3) given for the mean can be extended to the median as well. In principle the usual variance estimate s^2 can be used for σ^2. Also, one can set $\hat{\sigma}_m^2 = n\widehat{\mathrm{Var}}(\widehat{\mathsf{med}}(Y))$ and use any estimate for the variance of a median; see (2.8) for instance. So the main remaining limitation is the difference $E\widehat{\mathsf{med}}(Y) - E(Y_{n+1})$ – which can be large for highly skewed distributions. If there is enough data, one can draw bootstrap samples evaluate estimates of $E\widehat{\mathsf{med}}(Y)$ and $E(Y_{n+1})$ this for each sample and take the differnce. (This also gives an SE for the difference if desired.)

However, a technique that permits the distribution of Y_{n+1} to be asymmetric and does not involves the mean would be better. One choice is to use an estimate such as $\hat{S}$ for $S(\cdot)$ in the more general expressions (2.7) and (2.9). That is, rather than use a point predictor at all, form prediction intervals using the EDFs using the techniques of Sec. 2.1.2.

It is not hard to imagine that P is sufficiently uncertain as to make bounds of the form (5.7) not as good as desired. That is, they may be too narrow to be convincing or (more likely) too wide to be useful. After all, these methods assume the distribution F, or equivalently S, is stable enough and well enough defined to be taken as a constant. When this assumption fails, the natural way to proceed is to build variability due to P into the upper bound (5.7) by using a range of P's that might be valid models. Indeed, [39] explores two methods to account for model uncertainty in P. In one method, P is selected from a set candidate P's thereby increasing the uncertainty of downstream prediction. In the other method, the problem is enlarged so that P itself is a random outcome of an underlying process, again typically increasing the uncertainty of downstream prediction. Describing this is beyond our present scope.

For the sake of contrast, suppose our interest is not in the survival time itself but in the probability of survival to a given time, i.e., $\hat{S}(t)$. For example, it is common particularly in medical studies for there to be specific interest in the probability of survival of t years where $t = 1, 3, 5, 10$. Direct estimates of these probabilities are provided by the Kaplan-Meier estimate of the survival function, with confidence limits provided by a CI of the form in (5.6). If our interest is in the distribution of survival time Y_{n+1} of a new patient, conditional on the observed data, we can use the Kaplan-Meier estimate directly. This appears to work well when the sample size is large [20], but may lead to substantial cumulative error if prediction is for a group of individuals of size comparable to n. However, there are alternatives to the Kaplan-Meier estimator for this context, such as the predictive distribution proposed by Berliner and Hill [8].

The key issue, though, is that estimating the probability of survival to a given time is different from predicting an actual survival time and probability-based forecasting does not directly fit into the paradigm of Section 2.1. In fact, probability foreasting is on a different scale from predicting the next value of a random variable. If we only have the estimated probability of t year survival from a group of IID observations we cannot compare it with a new outcome without

further assumptions. We could invoke a criterion such as a scoring rule to enable comparison of outcomes with probabilities, see [37] for a recent contribution. This is not prequential because a scoring rule need not compare predictions with outcomes. Alternatively, we could reasonably compare a predicted probability with a collection of new outcomes that gave an estimated probability.

5.1.2. Discrimination and calibration

If our interest is prediction we may want to use the survival distribution as the basis of a diagnostic model, e.g., as a method for predicting a binary outcome for individuals, such as 'low-risk' and 'high-risk'. In essence, we want to take a *prognostic* model – one which predicts the *probability* of a future event or state – and convert it into a *diagnostic* model – one which predicts an actual future class or state. This means we have to define the future classes or states – often just called a category – and often there are many ways to do this. Once this is done, the accuracy of prognostic models is usually described in terms of two components, namely, discrimination and calibration [29]. *Discrimination* is the ability to classify individuals into their correct categories. *Calibration* is the ability to estimate the risk or probability of a future event accurately. Since our interest here is in the construction of predictors, not their evaluation, we only discuss discrimination briefly.

A survival distribution will provide predicted probabilities of survival, so we could convert this into a discriminant by classifying those with probabilities above a given cut point as belonging to one category and those with lower probabilities as belonging to another category. For such a case, we can use sensitivity and specificity, as introduced in Section 2.2.4, to choose the optimal cut point for accurate classification. One way to do cut point selection is via a receiver operating characteristic (ROC) curve [77]. This is a plot of sensitivity vs. (1-specificity) and presents all possible [sensitivity, (1-specificity)] pairs attainable by dichotomizing the probabilities of survival with different cutpoints. From the curve we can identify the threshold which maximizes both sensitivity and specificity, keeping in mind that the optimal threshold should also be a function of the relative costs of misclassifying subjects.

In many cases it is not clear which type of prognostic survival model will lead to the best diagnostic model. In addition to KM, we discuss accelerated failure time models and proportional hazards models in the next two subsections. Any of these models could be used to give a diagnostic model, however doing so is beyond our present scope.

5.2. Simple parametric families for survival data

One of the questions asked in prediction with lifetime data is conditional: Given that a subject has survived to time y, what is the probability of surviving longer? This is formally given by $P(Y \geq y + z|Y \geq y)$ where Y is interpreted

as a survival time. For a variety of reasons it is the conditional density that is used and called the 'hazard' function:

$$h(y) = \lim_{\delta \to 0} \frac{P(y \leq Y \leq y + \delta | Y \geq y)}{\delta} = \frac{p(y)}{S(y)} = -\frac{d \ln S(y)}{dy}.$$

The hazard is seen to be 'instantaneous' at y. So, integrating it gives the cumulative hazard function

$$H(y) = \int_0^y h(u)du = -\ln S(y), \tag{5.8}$$

and $S(y) = e^{-H(y)}$. The shape of the hazard function indicates how survival up to time y affects survival in the infinitesimal interval of time following y. It is easy to imagine survival up to y represents 'wearing out' so that $h(y)$ is increasing. It is also easy to imagine that survival up to y represents recovering from illness so that $h(y)$ is decreasing. All intermediate shapes are possible, too.

Given this framework it is possible to assign parametric families to P and see what the corresponding hazard function looks like. One choice for P is the Exponential(λ) distribution. It can be verified that $h(y) = \lambda$, i.e., is constant, and that

$$P(Y \geq y + z | Y \geq y) = P(Y \geq z)$$

i.e., the Exponential(λ) is memoryless in the sense that having waited y units and then waiting a further z units is the same as having waited for z units in the first place.

A generalization of the Exponential is the Weibull(α, λ), with survival function $S(y) = e^{-\lambda y^\alpha}$. The extra parameter α affects the shape of the underlying density and the hazard function is $h(y) = \lambda \alpha x^{\alpha - 1}$, increasing for $\alpha > 1$, decreasing for $\alpha < 1$, and constant for $\alpha = 1$. There are numerous other distributions that have been studied in the context of lifetime data including the Gamma, the log-normal, the extreme value, and the Pareto.

Whatever the parametric family chosen, the procedure for prediction of survival time is much the same: Form a predictor, whether a mean, median, mode or other sort of statistic, and control the probability that it is a given distance away from the next outcome. This is conceptually the same IID prediction scenario as seen in Sec. 2.1 although in many cases parameters will have to be estimated and their variability assessed analogous to (2.1) and (2.3). If closed forms for predictive procedures do not exist, we may have to plug in estimates of parameters, possibly using upper or lower confidence bounds to get conservative bounds in PIs. If our interest is in probabilities of survival, these can be obtained directly from the estimated survival function (and its confidence bounds) as in Section 5.1.1.

One step up in complexity from either using the Kaplan-Meier or just assigning a plausible parametric model is the use of explanatory variables to estimate the survival function. The idea is that the explanatory variables should enable one to get tighter bounds on S or at least get tight enough bounds that interesting questions on the difference between groups in terms of their survival characteristics can be answered.

An important class of models for this scenario is called accelerated failure time (AFT) models in which failure times Y are accelerated by a factor depending on a d-dimensional covariate x, a parameter β and a specified link function g. The general parametric case is any specification of the conditional survival function $S(y|x)$ with the property that

$$S(y|x) = S_0(yg(x^T\beta)) \tag{5.9}$$

where S depends on β, S_0 is an unspecified baseline survival function, and x may be an outcome from a random variable X, see [50]. Both β and S_0 must be estimated to use (5.9).

The most commonly used example of an AFT model supposes

$$\ln Y = x^T\beta + \sigma\epsilon, \tag{5.10}$$

i.e., the exponential link function $g(u) = e^{-u}$ is adopted and the log of the survival time follows a linear model; the error term ϵ is often chosen to be a $N(0,\sigma^2)$, extreme value or logistic distribution even though (5.9) does not require additive error. It is seen that (5.10) gives

$$Y = e^{x^T\beta}e^{\sigma\epsilon} \tag{5.11}$$

so when ϵ is normal, (5.11) gives a log-normal regression model; when ϵ is extreme value (5.11) gives a Weibull regression model; and when ϵ is logistic, (5.11) gives a log-logistic regression model. Each of these choices of error term leads to a model class and the classes are studied individually; see [55, 7] among others. See [50] for a case in which S_0 is modeled as a mixture of parametric survival functions in which the mixing distribution must be specified. Using an exponential link, [50] adopts a hierarchical Bayesian model giving an analog to (5.10), see equation (3); the survival time is taken as Weibull.

Clearly, (5.10) implies the AFT property (5.9):

$$\begin{aligned} S_{\beta,x}(y) &= P_\beta(Y > y|x) = P_\beta(\sigma\epsilon > \ln y - x^T\beta|x) \\ &= P_\beta(e^{\sigma\epsilon} > ye^{-x^T\beta}|x) = S_0(ye^{-x^T\beta}), \end{aligned}$$

so survival time is accelerated, relative to S_0, by $e^{-x^T\beta}$ the link function evaluated at the regression function. Other link functions give analogs of (5.10) and (5.12). If h_0 is a baseline hazard function corresponding to S_0, i.e., from $x = 0$, then an acceleration is seen on the hazard scale as well as the survival function scale. For the exponential link this is $h_{\beta,x}(y) = h_0(ye^{-x\beta})e^{-x\beta}$ and other link functions give analogous results. So, again, the effect of the covariates is to affect the shape of a baseline hazard rate in effect increasing or decreasing the chance for survival given survival up to a fixed time.

Given data of the form (x_i, Y_i) for $i = 1,\ldots,n$ and an x_{n+1} for which we would like to predict $Y_{n+1}(x_{n+1})$, denote the predictor by $\hat{y}(x_{n+1})$. As an example, if we fix a $N(0,\sigma^2)$ error distribution so that the variability of the future outcomes could be determined once the parameters were known and estimated

β and σ by $\hat{\beta}$ and $\hat{\sigma}$, for instance by maximum likelihood, then we could form the survivor function

$$S_{\hat{\beta},x_{n+1}}(y) = P_{\hat{\beta}}(\hat{\sigma}\epsilon > \ln y - x_{n+1}\hat{\beta}|x_{n+1}). \tag{5.12}$$

Converting (5.12) to an estimator $\hat{F}$ of $F_{\beta,x}$ we can obtain a PI as in (2.7) or (2.9) and take the midpoint as a point predictor. Alternatively, we could use the median survival time from $S_{\hat{\beta},x_{n+1}}(y)$ as a point predictor, investigating its properties as in (2.1) and (2.3). We can also get point predictions of t year survival directly from $S_{\hat{\beta},x_{n+1}}(y)$. However, none of these approaches is really satisfactory because the variability in $\hat{\beta}$ and $\hat{\sigma}$ is neglected. Including it would typically make the prediction intervals larger. On the other hand, for some predictive purposes, ignoring the variability in $\hat{\beta}$ (or $\hat{\sigma}$) may not be too damaging. Indeed, [7] develops deletion diagnostics for influential data points to stabilize percentile based predictions from AFT models with exponential link functions; this may be more important than the variability in $\hat{\beta}$ per se.

5.3. Proportional hazards and prediction

This strategy for finding a good predictor of survival starts by modeling the hazard function rather than the survivor function. Then, one converts the estimate of the hazard function into an estimate for the survivor function so as to make predictions. One model for the hazard function given x is

$$h(y|x) = \psi(x)h_0(y) \tag{5.13}$$

where ψ is a function of the covariates that relates the baseline hazard function to the hazard function for non-zero covariates. Often, ψ is taken to have a parametric form in which the information in x is summarized by a linear function. In these cases, (5.13) can be written

$$h(y|x) = h_\beta(y|x) = \psi(x\beta)h_0(y). \tag{5.14}$$

The most common form for ψ in (5.14) is an exponential so that $\psi(x\beta) = e^{x\beta}$. This makes ratios of hazard functions independent of y. That is, $\ln \frac{h(y|x)}{h_0(y)} = x\beta$ and this is described by saying the hazards are proportional. In fact, it is easy to see that the log-hazard ratio is linear, i.e., for any covariates x and x',

$$\ln \frac{h(y|x)}{h(y|x')} = (x - x')\beta.$$

The foundational analysis of proportional hazards (PH) models is in [30] for which reason they are often called Cox models. To obtain predictions from a PH model, one must find estimates $\hat{\beta}$ of β and $\hat{h}_0$ of h_0. The need for β is obvious, but prediction requires the use of the baseline hazard also since predictions rely on the distribution function and cannot be made from quantities that are

proportions in which h_0 might cancel. So, suppose that we have data of the form (y_i, δ_i, x_i) for $i = 1, \ldots, n$, where x_i are the covariates for subject i and subject i is observed for time y_i. Each δ_i is an indicator function for whether the time y_i is right censored ($\delta_i = 0$) or whether y_i really is an observation of Y ($\delta_i = 1$), i.e., whether y_i is just the end of the observation period with no event observed or we actually observed an event at time y_i. Since we may regard the x_i as outcomes of a random variable X_i, we assume as needed that δ_i and X_i are independent. This means that the censoring process and the explanatory variables are unrelated.

The complete likelihood under right-censoring with exponential ψ is derived concisely in [62] (p. 75-76, 258) as

$$\begin{aligned} L(\beta, h_0(\cdot)) &= \prod_{i=1}^{n} h_\beta(y_i|x_i)^{\delta_i} S_\beta(y_i|x_i) \\ &= \prod_{i=1}^{n} h_0(y_i)^{\delta_i} (e^{x_i^T\beta})^{\delta_i} e^{-H_0(y_i)e^{x_i^T\beta}}, \end{aligned} \tag{5.15}$$

using (5.8) for the second equality. In fact, although $L(\beta, h_0)$ appears to depend on the whole function h_0, it only depends on the specific values $h_0(y_i)$.

So, it is enough to maximize (5.15) over the $h_0(y_i)$'s (and β). In general, there will be D 'deaths', i.e., values of y_i that are observed without censoring, $D \leq n$. So, (5.15) can be simplified by separating the D cases with $\delta_i = 1$ from the $n - D$ cases with $\delta_i = 0$. Relabeling the y_i's as required in the first factor, this gives

$$\begin{aligned} L(\beta, h_0(y_1), \ldots, h_0(y_D)) &= \left[\prod_{i=1}^{D} h_0(y_i) e^{x_i^T\beta}\right] e^{-\sum_{i=1}^{n} H_0(y_i)e^{x_i\beta}} \\ &\propto \prod_{i=1}^{D} h_0(y_i) e^{-h_0(y_i)\sum_{j\in R(y_i)} e^{x_j^T\beta}} \end{aligned} \tag{5.16}$$

in which the risk set $R(y)$ for any time y is the set of subjects who are 'at risk' at time y. (The derivation giving the proportionality in (5.16) is not obvious, but not hard either.) Thus, $R(y_i)$ groups together the set of individuals who are alive and uncensored just prior to y_i. Expression (5.16) is a partial likelihood since some factors in β are dropped.

Maximizing (5.16) over the $h_0(y_i)$'s gives

$$\hat{h}_0(y_i) = \hat{h}_{0,\beta}(y_i) = \frac{1}{\sum_{j\in R(y_i)} e^{x_j^T\beta}} \tag{5.17}$$

and $\hat{H}_{0,\beta}(y) = \sum_{y_i \leq y} [1/\sum_{j\in R(y_i)} e^{x_j^T\beta}]$.

It remains to find $\hat{\beta}$. To express this, write $x_{(i)}$ to mean the covariate value x_i associated with the i-th subject having time of death or last observation y_i.

Putting (5.17) into (5.16) for $h_0(y_i)$ and putting back the factors left out in (5.16) gives

$$L^*(\beta) = \prod_{i=1}^{t} \frac{e^{x_{(i)}^T \beta}}{\sum_{j \in R(y_i)} e^{x_{(i)}^T \beta}}, \tag{5.18}$$

see [30] or [62]. [30] p. 191 calls L^* a conditional likelihood and [62] p. 258 refers to L^* as a profile likelihood; [60] p. 269 calls L^* a marginal likelihood.

The log of (5.18) can be maximized over β by a variety of iterative methods, including Newton-Raphson. The Fisher information matrix can also be obtained and evaluated at $\hat{\beta}$ to obtain an estimate of the variance matrix of $\hat{\beta}$ by inversion.

Given $\hat{\beta}$, the estimate $\hat{h}_0(y_i)$ from (5.17) is complete. So, (5.14) can be used to get an estimate of $h(y|x)$ (since ψ is assumed known). Using $\hat{h}_0(y_i)$ we can also obtain $\hat{H}_0(y) = \hat{H}_{0,\hat{\beta}}(y)$ for any y and hence $\hat{H}_{\hat{\beta},X}(y)$ for any y. In addition, (5.8) provides an estimate $\hat{S}_0(y) = \hat{S}_{0,\hat{\beta}}$ of $S_0(y)$ and hence $\hat{S}_{\hat{\beta},x}(y)$ for any covariates x and y. We comment that a different approach to this optimization is pursued in [60] Sec. 4, Eq. (7) and (8). It leads to a different estimate of $\hat{h}_0$, $\hat{H}_0$ and hence $\hat{S}_0$ and hence to different estimates $\hat{h}_{\hat{\beta},x}$, $\hat{H}_{\hat{\beta},x}$ and hence $\hat{S}_{\hat{\beta},x}$ when covariates are included.

As before, once the survival function $\hat{S}_{\hat{\beta},x}(y)$ is obtained, PI's for any value of x can be given by taking percentiles from $P_{\hat{\beta},\hat{h}_0,x}$, neglecting variability in $\hat{\beta}$ and $\hat{h}_0$. These intervals may also be obtained, parallel to (2.7) or (2.9) by invoking an asymptotic normality argument, see [69]); the median i.e., 50th percentile would be a natural choice for a point predictor. Alternatively, PI's from a given point predictor, such as a mean or median, can be found by using $P_{\hat{\beta},\hat{h}_0,x}$ in (2.1) or (5.7) if the variability in $\hat{\beta}$ and $\hat{h}_0$ is neglected. More carefully, one would have to incorporate the variability in $\hat{\beta}$ and $\hat{h}_0$, analogous to (2.3).

We conclude this section by noting that as a predictive strategy, PH models have their detractors. Indeed, it is not clear that survival analysis can, as commonly used, provide more than a serviceable summary of a data set. Specifically, predictions seem, often, to be too weak to be useful; see [53] and [54]. This may be due to modeling being poor (e.g., the hazards are only approximately proportional) or difficult (e.g., the hazard function is too complex compared to the models used to approximate it) or to the high intrinsic variability of the biomedical populations to which PH models are most frequently applied. In any of these case, PH models may require more effective validation than is commonly done. One further possibility is that the relationship between data and a conditional density such as a hazard function is much more distant than the relationship between data and its distribution function. This may mean that reliably estimating a hazard function for high intrinsic variability populations just requires more data than are typically available.

6. Summary

It can be seen from Sections 2 and 3 that predictive techniques for basic independent data, e.g., as occurs in linear regression and elementary classification, and time series data are well developed. Various predictive techniques for longitudinal data, Section 4, are also well developed but the extra variability from modeling the population – especially the random effects – will limit their effectiveness. In addition, predictions from generalizations of linear mixed models (such as GLMS's, GEE's, GLMM's, and NLMM's) do not seem to have been explored and in some cases such as GEE's are conceptually problematic. In some cases, however, prediction can be done effectively with large enough sample sizes – although it is unclear when the PI's from, say, GLM models, will be sufficiently narrow as to be useful. For survival analysis, we have described the obvious predictive techniques in Sec. 5. As with longitudinal analysis, predictive techniques are not as commonly used with survival data as one might expect or hope. But, unlike longitudinal analysis, it is not clear why. It may be that the goal of survival analysis is so oriented to modeling that prediction is ignored or it may be that model uncertainty and the intrinsic variability of the populations being modeled is so often high relative to the sample sizes commonly obtained that predictions are unreliable or PI's too wide to be useful.

Indeed, one of the problems with prediction is that point predictors more variable than point estimators and PI's are typically wider than CI's. Moreover, just like CI's, model-based PI's tend to enlarge when model uncertainty is taken into account. The consequence of this is that predictive inferences tend to be weaker than parametric or other inferences about model classes. It would be natural for investigators to prefer stronger statements – even if the justification for them rests heavily on ignoring model uncertainty. However, even though inferentially weaker, point predictors and PI's have the benefit of direct testability that point estimators and CI's usully lack.

It must also be admitted that the predictive approach is frequently harder than modeling: It's easier in general to find a not-implausible model, estimate a few parameters, verify the fit is not too bad and then use the model to make statements about the population as a whole than it is to find a model that is not just plausible but actually close enough to correct to give good predictions for new individual members of the population. Here, 'close enough' means that the errors from model mis-specification or model uncertainty are small enough, compared with those from other sources of error, that they can be ignored. Unfortunately, however, it seems that there are so many plausible models that finite data sets often cannot discriminate effectively amongst them. That is, as a generality, the plausibility of a model is insufficient for good prediction because one is quite likely to have found an incorrect model that the data have not been able to rule out. Since models that do not give sufficiently good prediction have to be disqualified, their suitability for other inferential goals must be justified by some argument other than goodness of fit.

The effect of model uncertainty is explored in [39] who compares two ways of accounting for model uncertainty in post-model selection inference, including

prediction. [39] argues that model enlargement – basically adding an extra level to a Bayesian hierarchical model – is a better solution than trying to account for the variability of model selection from criteria such as AIC and BIC. He also argues that it is better to tolerate larger prediction intervals than to model uncertainty incorrectly. (As a curious note, [40] finds that there are cases where correctly accounting for modeling uncertainty actually reduces predictive uncertainty.) Of course, if PI's are too large to be useful then the arguments that a modeling approach is valid are more difficult to make and any other inferences – estimates, hypothesis tests – may be called into question. While we have not focussed on model uncertainty, we have mentioned concerns arising from it in various places including Secs. 4.2 and 5.1.1 and the end of Sec. 5.3.

A separate issue from model uncertainty is model accessibility i.e., the degree to which a data generator can be represented as an identifiable model. Aside from the possibility of representing a 'true model' probabilistically (e.g., the true model for Y_i is $p_\alpha(\cdot|\theta_\alpha)$ where $\alpha \sim w(\alpha)$ and different Y_i's have different α_i's), it may be that the true model is effectively inaccessible in the sense of not being expressible in any useful form. Even more, it may be, for some data generators, that the true model does not even exist in a meaningful sense. These are the cases of $\mathcal{M}$-complete and $\mathcal{M}$-open, see [5]. All the techniques in this paper are intended for the $\mathcal{M}$-closed case and it is not clear how well they extend to $\mathcal{M}$-complete and $\mathcal{M}$-open problems – even though $\mathcal{M}$-complete and $\mathcal{M}$-open problems may be more typical of the subject matter for which the techniques in this paper are intended.

Nevertheless, the main strategy has been to look at model classes and use them to generate predictors. However, the reverse may be more reasonable especially for complex or high dimensional data. That is, one may propose a predictor class and find a member that performs well. Then, if model identification is desireable for some reason, one can, in principle, convert the predictor to a model within a class of models that are believed plausible. For instance, in some settings Bayes model averaging yields a good predictor. One can form a single model from it by looking at the leading terms in the models that went into the average. As another example, one can use a kernelized method such as a relevance vector machine (RVM), take a Taylor expansion of the kernel in each term of the RVM and again take the leading terms as a model. In this way one might obtain a model that is at interpretable and gives good predictions, even if the predictions are not quite as good as those from the original predictor.

There are numerous properties good predictors should have that have not been discussed here. For instance, a good predictor should be stable to perturbations of the data set that went into forming it. One common way to establish stability is to perturb the data with, say, normal noise (possibly allowing the mean to be nonzero), and then rerunning the procedure by which the predictor was generated to verify the new predictor is not unacceptably far from the original predictor. Other considerations and constraints on predictors are also important, see [25] for a general discussion.

Finally, we comment that probability forecasting has been advocated by numerous authors. We have not done this – limiting ourselves to brief comments in

Sec. 5.1.1 and 5.1.2 – because our goal has been prediction of outcomes directly. In some settings, probability forecasting may be more informative, in the way that soft classifiers may be more useful in many settings than hard classifiers. However, probability forecasts involve quantities such as probabilities that are not directly measureable and here we have limited our attention to quantities that are on the same scale as directly measureable quantities.

Appendix A: Dynamic linear models

The basic idea behind the structure of dynamic linear models (DLM) is to make just enough assumptions that sequential prediction is feasible. A good introductory treatment can be found in [79] and a more detailed treatment can be found in [97]. The treatment here is a partial summary of their presentation focusing on the predictive structure in the univariate case.

The overall DLM structure is expressed in two equations. The first is called the observation equation,

$$Y_i = x_i^T \beta_i + \epsilon_i, \tag{A.1}$$

and the second is called the system equation,

$$\beta_i = G^T \beta_{i-1} + u_i. \tag{A.2}$$

It is only the Y_i in (A.1) that is observed with error ϵ_i; the x_i is known and presumed constant in an analogy with a design matrix. The relationship between Y_i and x_i is controlled by β_i and (A.2) means that β_i follows an autoregressive model in which the matrix G is taken as known. (The matrix G may be dependent on i but we ignore this case here.) This means that the time evolution of the coefficients β_i is deterministic apart from the error u_i. At least for values of i and i' that are close together, β_i and $\beta_{i'}$ are not far apart, but of course as $i - i'$ increases, the corresponding β_i and $\beta_{i'}$ may end up far apart. Therefore, as $i - i'$ increases, the information in Y_i about $Y_{i'}$ decreases. This is a way to build some variability into (A.1) apart from the ϵ_i. In the simplest cases, $\epsilon_i \sim N(0, V_i)$ and $u_i \sim N(0, W_i)$ and all are assumed to be independent.

To see how updating works in the DLM, suppose that

$$\beta_{i+1}|D_i \sim N(\mu_{i+1}, C_{i+1}), \tag{A.3}$$

where D_i represents the knowledge available at time i. That is, the information in D_{i-1} is 'contained' somehow in the information in D_i. We must assume an initial distribution for $(\beta_0|D_0)$, say a normal, but at this stage, D_i need not be specified further. If a squared error loss is assumed so that conditional expectations are optimal predictors then

$$E(Y_{i+1}|D_i) = x_{i+1}^T \mu_{i+1}, \tag{A.4}$$

and using the independence relations among the error terms we get

$$\mathrm{Var}(Y_{i+1}|D_i) = x_{i+1}^T C_{i+1} x_{i+1} + V_{i+1}, \tag{A.5}$$

so that $(Y_{i+1}|D_i) \sim N(X_{i+1}^T \mu_{i+1}, x_{i+1}^T C_{i+1} x_{i+1} + V_{i+1})$. Forecasting two steps into the future requires an extra use of the system equation and in general forecasting ℓ steps into the future requires ℓ uses of the system equation to update the distribution on the β_i to $\beta_{i+\ell}$. The result is

$$\beta_{i+\ell}|D_i \sim N(\mu_i(\ell), C_i(\ell)), \tag{A.6}$$

in which

$$\mu_i(\ell) = G^{\ell-1}\mu_{i+1}$$

and

$$C_i(\ell) = G^{\ell-1}C_{i+1}(G^{\ell-1})^T + \sum_{j=2}^{\ell} G^{\ell-j}W_{i+j}(G^{\ell-j})^T.$$

So, predictions are obtained from the observation equation for ℓ steps,

$$Y_{i+\ell}|D_i \sim N(X_{i+\ell}^T \mu_i(\ell), F_{i+\ell}^T C_i(\ell) F_{i+\ell} + V_{i+\ell}). \tag{A.7}$$

Analogous derivations for sums such as $Z_{i+1}^{i+k} = \sum_{j=i+1}^{i+k} Y_j$ give $E(Z_{i+1}^{i+k}|D_i)$ and $\mathrm{Var}(Z_{i+1}^{i+k})$.

To complete the picture, note that the system equation gives the likelihood

$$L(\beta_i|Y_i = y_i, V_i) \propto p(Y_i = y_i|\beta_i, V_i) \sim N(x_i^T \beta_i, V_i)$$

and using this the posterior for θ_i is given by

$$p(\beta_i|D_{i-1}, y_i) = \frac{p(Y_i = y_i|\beta_i, V_i)p(\beta_i|D_{i-1})}{p(Y_i = y_i)}. \tag{A.8}$$

Since all three densities on the right hand side are normal, $(\beta_i|D_{i-1}, y_i)$ is also normal and the prior on β_i can be updated to give updated forecasts for $Y_{i+\ell}$. In addition, $(\beta_{i+\ell}|D_i)$ can also be derived to give an updated prior for $\beta_{i+\ell}$ in place of that in (A.8). See Chapter 4 of [97] for full details on the derivations.

Acknowledgements

J. Clarke's research was partially funded by National Institutes of Health (NIH) grant 5K25CA111636-11.

The authors express their gratitude to the Associate Editor and two anonymous referees whose comments greatly improved the exposition. Indeed, one referee gave us over 20 pages (!) of help and insight. We hope he or she will derive some satisfaction from how thoroughly we have acted on those comments.

References

[1] Agresti, A. (2002) *Categorical data analysis* 2nd Ed. Wiley and Sons, New York. MR1914507

[2] Aitchison, Y. (1975). Goodness of prediction fit. *Biometrika* **62** 547–554. MR0391353

[3] Altman, D. and Bland, J. (1994). Diagnostic tests. 1: Sensitivity and Specificity. *British Medical Journal* **308** 1552.

[4] Bannerjee, S. (2008) Bayesian Linear Models: The Gory Details. *Dowloaded from* http://www.biostat.umn.edu/~ph7440/

[5] Bernardo, J. M. and Smith, A. F. M. (2000) *Bayesian Theory*. John Wiley and Sons, Chichester. MR1274699

[6] Barnett, G., Kohn, R. and Sheather, S. (1997) Robust Bayesian estimation of autoregressive-moving-average models. *J. Time Series* **18**, 11–28. MR1437739

[7] Bedrick, E., Exuzides, A. Johnson, W. and Thurmond, M. (2002) Predictive influence in the accelerated failure time model. *Biostatistics*, **3**, 331–346.

[8] Berliner, L., and Hill, B. (1988) Bayesian nonparametric survival analysis. *J. Amer. Stat. Assoc.*, **83**, 772–779. MR0963805

[9] Booth, J. and Hobert, J. (1998) Standard errors of prediction in GLMM's *J. Amer. Stat. Assoc.*, **93**, 262–272. MR1614632

[10] Booth, J. and Hobert, J. (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm *J. Roy. Stat. Soc. Ser. B*, **61**, 265–285.

[11] Box, G. and Jenkins, G. (1970) *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco. MR0272138

[12] Box, G. and Jenkins, G. (1976) *Time Series Analysis: Forecasting and Control*. Revised Edition, Holden-Day, San Francisco. MR0436499

[13] Breslow, N. and Crowley, J. (1974) A large sample study of the life table and product-limit estimates under random censorship. *Ann. Stat.*, **2**, 437–453. MR0458674

[14] Brockwell, P. and Davis, R. (1987) *Time Series: Theory and Methods*. Springer, New York. MR0868859

[15] Burnham, K., and Anderson, D. (2002) *Model selection and multimodel inference: a practical information-theoretic approach.* 2nd Ed. Springer-Verlag, New York. MR1919620

[16] Butler, R. (1986) Predictive likelihood inference with applications. *J. Roy. Stat. Soc. Ser. B*, **48**, 1–38. MR0848048

[17] Carpenter, S. (2003). *Regime Changes in Lake Ecosystems: Pattern and Variation*. Volume 15 in the Excellence in Ecology Series, Ecology Institute, Oldendorf/Luhe, Germany.

[18] Chandler, R. and Scott, E. (2010) *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences* John Wiley and Sons, London. MR2829966

[19] Chandler, R. and Skouras, R. (1998) *Forecasting Class Notes*. Dept. Statistical Science, UCL, London.
[20] Chang, M. and Shuster, J. (1994) Interim Analysis for randomized clinical trials: Simulating the log-rank test statistic. *Biometrics*, **50**, 827–833.
[21] Chen, K. and Lo, S. (1997) On the rate of uniform convergence of the product limit estimator: weak and strong laws. *Ann. Stat*, **25**, 1050–1087. MR1447741
[22] Chi, E. and Reinsel, G. (1989) Models for longitudinal data with random effects and AR(1) errors. *J. Amer. Statist. Assoc.* **84**, 452–459. MR1010333
[23] Chib, S. and Greenberg, E. (1994) Bayes inference in regression models with $ARMA(p,q)$ errors. *J. Econometrics* **64**, 183–206. MR1310523
[24] Christensen, R., Johnson, W., Branscum, A. and Hanson, T. (2011) *Bayesian Ideas and Data Analysis*. Chapman & Hall/CRC Press. MR2682928
[25] Clarke, B. (2010). Desiderata for a predictive theory of statistics. *Bayes Analysis*, **5**, 283–318. MR2719654
[26] Clarke, B., Fokoue, E. and Zhang, H. (2009) *Principles and theory for Data Mining and Machine Learning*. Springer, New York. MR2839778
[27] Cole, T. and Green, P. (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Stat. Med.* **11**, 1305–1319.
[28] Collett, D. (1994) *Modelling survival data in medical research*. Chapman & Hall, London.
[29] Cook, N. (2008) Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin. Chem.* **54**, 17–23.
[30] Cox, D. (1972) Regression models and life tables (with discussion). *J. Roy. Stat. Soc. Ser. B* **74**, 187–220. MR0341758
[31] DasGupta, A. (2008) *Asymptotic Theory of Statistics*. Springer, NY. MR2664452
[32] Davidian, M. and Giltinan, D. (2003) Nonlinear models for repeated measurements: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics* **8**, 387–419.
[33] Davis, C. (2002) *Statistical methods for the analysis of repeated measurements*. Springer, New York. MR1883764
[34] Davis, R. and Dunsmuis, T. (1997) Least absolute deviation estimation for regression with $ARMA$ errors. *J. Theor. Probab.* **10**, 481–497. MR1455154
[35] Dawid, A. P. and Vovk, V. (1984). Prequential probability: Principles and properties. *Bernoulli* **5** 125–162. MR1673572
[36] Dawid, A. P. (1984). Statistical theory: The prequential approach. With Discussion. *J. Roy. Statist. Soc. A* **147**, 278–292. MR0763811
[37] Dawid, A. P., Lauritzen, S., and Parry, M. Proper Local Scoring Rules on Discrete Sample Spaces. *Ann. Statist.*, to appear.
[38] de Gooijer, J., Abraham, B., Goulde, A., and Robinson L. (1985). Methods for determining the order of an autoregressive-moving aver-

age process: A survey. *International Statistical Review A* **53**, 301–329. MR0967215

[39] Draper, D. (1995) Assessment and propagation of model uncertainty. *J. Roy. Stat. Soc.*Ser. B, **57**, 45–97. MR1325378

[40] Draper, D. (1997) On the relationship between model uncertainty and inferenetial/predictive uncertainty. *Unpublished manuscript.*

[41] Diggle, P., Liang, K., and Zeger, S. (1996) *Analysis of Longitudinal Data* Oxford University Press, Oxford.

[42] van Erven, T., Grunwald, P., de Rooij, S. (2012) Catching up faster by switching sooner: A prequential solution to the AIC-BIC dilemma. *J. Roy. Stat. Soc.* Ser. B, to appear.

[43] Feller, (1948) On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions. *Ann. Math. Stat.* **19** (2), 177–189. MR0025108

[44] Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**, 179–188.

[45] Fitzmaurice, G., Laird, N. and Ware, J. (2004) *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics, New York. MR2063401

[46] Foldes, A. and Rejto, L. (1981) Strong uniform consistency for nonparametric survival curve estimators from randomly censored data. Ann. Stat., 9, 122–129. MR0600537

[47] Franses, P. (2002) Testing for residual autocorrelation in growth curve models. *Tech. Forecasting and Soc. Change* **69**, 195–204.

[48] Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis* 2nd Ed. Chapman & Hall/CRC Texts in Statistical Science. MR1385925

[49] Geisser, S. (1995). *Predictive Inference: An Introduction*. Chapman & Hall, New York, NY. MR1252174

[50] Ghosh, S. and Ghosal, S. (2006) Semiparametric accelerated failure time models for censored data. Chapter 15 in: *Bayesian Statistics and Its Applications*, Upadhyay, S., Singh, U., and Dey, D. (Eds.) Anamaya Publishers, New Delhi.

[51] Gill, R. (1983) Large sample behavior of the product limit estimator on the whole line. *Ann. Stat.*, **11**, 49–58. MR0684862

[52] Hannan, E. (1980) Estimation of the order of an ARMA process. *Ann. Statist.* **8**, 1071–1081. MR0585705

[53] Henderson, R., Jones, M. and Stare, J. (2001) Accuracy of point predictions in survival analysis. *Stat. Med.* **20**, 3083–3096.

[54] Henderson, R. and Keiting, N. (2005) Individual survival time prediction using statistical models. *J. Med. Ethics.* **31**, 703–706.

[55] Hosmer, D. and Lemeshow, S. (1999) *Applied survival analysis*. Wiley, New York. MR1674644

[56] Ibrahim, J., Chen, M.-H., and Sinha, D. (2001) *Bayesian survival analysis*. Springer, New york. MR1876598

[57] James, G., Hastie, T. and Sugar, C. (2000) Principal component models for sparse functional data. *Biometrika*, **87**, 587–602. MR1789811

[58] Jayawardhana, A., and Samaranayake, V. (2004) Prediction bounds for the Weibull distribution. http://interstat.statjournals.net/YEAR/2004/abstracts/0411002.php
[59] *Linear and Generalized Linear Mixed Models and their Applications*. Springer, New York.
[60] Kalbfleisch, J. and Prentice, R. (1973) Marginal likelihoods based on Cox's regression and life model. *Biometrika*, **60**, 267–278. MR0326939
[61] Kaplan, E. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 457–481. MR0093867
[62] Klein, J., and Moeschberger, M. (2003) *Survival Analysis*. Springer, New York.
[63] Koenker, R. (2005) *Quantile Regression*. Campridge University Press, Cambridge. MR2268657
[64] Koenker, R. and Hallock, K. (2001) Quantile regression. *J. Econ. Perspectives* bf15, 143–156.
[65] Koenker, R. and Basset, G. (1978) Regression quantiles. *Econometrika*, **46**, 33–50. MR0474644
[66] Laird, N. and Ware, J. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.
[67] Lejeune, M., and Faulkenberry, G. (1982). A simple predictive density function. *J. Amer. Statist. Assoc.* **77**, 654–657. MR0675894
[68] Lindstrom, M. and Bates, D. (1990) Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673–687. MR1085815
[69] Link, C. (1984). Confidence intervals for the survival function using Cox's proportional-hazard model with covariates. *Biometrics* **40**, 601–610. MR0775377
[70] Meinshausen, N. (2006) Quantile regression forests. *J. Mach. Learn. Res.* **7**, 983–999. MR2274394
[71] Meyer, M. and Laud, P. (2002) Predictive variable selection in generalized linear models. *J. Amer. Stat. Assoc.*, **97**, 859–871. MR1941415
[72] Monahan, J. (1980) A structured approach to ARMA time series models, Part I: Distributional results. *Inst. Stat. Mimeo Series, #1297* NC State Univ. Raleigh.
[73] Monahan, J. (1983) Fully Bayesian analysis of ARMA time series models. *J. Econometrics* **21**, 307–331.
[74] Pan, J. and Fang, K. (2002) *Growth Curve Models*. Springer, New York.
[75] Pan, W. and Le, C. (2001) Bootstrap model selection in GLM's. *Biometrics*, **6**, 49–61. MR1812091
[76] Panhard, X. and Samson, A. (2009) Extension of the SAEM algorithm for nonlinear mixed models with 2 levels of random effects. *Biostatistics* **10**, 121–135.
[77] Pepe, M. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford, UK. MR2260483
[78] Philippe, A. (2006) Bayesian analysis of autoregressive moving average processes with unknown orders. *Comp. Stat. Data Anal.* **51**, 1904–1923. MR2307551

[79] Pole, A., West, M. and Harrison, J. (1994) *Applied Bayesian Forecasting and Time Series Analysis*. Chapman and Hall, New York.
[80] Proust-Lima, C. and Taylor, J. (2009) Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post-treatment PSA: a joint modeling approach. *Biostatistics*, **10**, 535–549.
[81] Rao, C. R. (1987). Prediction of future observations in growth curve models (with discussion) *Stat. Sci.* **2** 434–447. MR0933738
[82] Reiss, R.-D. (1989) *Approximate Distributions of Order Statistics*. Springer, NY. MR0988164
[83] Rice, J. and Wu, C. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–259. MR1833314
[84] Rissanen, J. (1996). Fisher information and stochastic complexity *IEEE Trans. Inform. Theory* **42** 40–47. MR1375327
[85] Robinson, G. K. (1991) That BLUP is a good thing: The estimation of random effects. *Stat.Sci.*, **6**, 15–32. MR1108815
[86] Sakia, R. (1992) The Box-Cox transformation technique: A review. *The Statistician*, **41**, 169–178.
[87] Shi, M., Weiss, R. and Taylor, J. (1996) An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *J. Roy. Stat. Soc.* Ser. C **45**, 151–163.
[88] Steyerberg, E., Vickers, A., Cook, N., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M., and Kattan, M. (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 139–141.
[89] Taylor, J. (2000) A Quantile Regression Neural Network Approach to Estimating the Conditional Density of Multiperiod Returns. *J. Forecasting* **19**, 299–311.
[90] Therneau, T. and Grambsch, P. (2000) Modeling Survival Data: Extending the Cox Model. Springer, New York. MR1774977
[91] Thomas, W. and Cook, R. D. (1990) Assessing influence on predictions from generalized linear models. *Technometrics*, **21**, 59–65. MR1050280
[92] Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, **92**, 937–950. MR2234196
[93] Yu, K. and Moyeed, R. (2001) Bayesian quantile regression *Stat. Prob. Letters* **54**, 437–447. MR1861390
[94] Verbeke, G. and Molenberghs, G. (2009) Linear Mixed Models for Longitudinal Data Springer, New York. MR2723365
[95] Waller, L. and Gotway, C. (2004) *Applied spatial statistics for public health data* John Wiley and Sons, New York. MR2075123
[96] Wong, H. and Clarke, B. (2004) Improvement over Bayes prediction in small samples in the presence of model uncertainty. *Can. J. Stat*, **32**, 269–283. MR2101756
[97] West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Linear Models* Springer, New York. MR1482232

[98] Zellner, A. (1971) *An Introduction to Bayesian Analysis in Econometrics* Wiley, New york. MR0433791

[99] Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Goel, P. and Zellner, A. (Eds.) p. 233–243, North Holland/Elsevier, Amsterdam. MR0881437

[100] Zellner, A. and Geisel, M. (1970) Analysis of distributed lag models with applications to consumption function estimation. *Econometrika* **38**, 865–888.

[101] Zellner, A. and Williams, A. (1973) Bayesian analysis of the Federal Reserve-MIT-Penn model's Almon lag consumption function. *J. Econometrics* **1**, 267–299.

[102] Zhang, H. (2002) On estimation and prediction for spatial generalized linear mixed models. *Biometrics* **58**, 129–136. MR1891051

[103] Zinde-Walsh, V. and Galbraith, J. (1991) Estimation of a linear regression model with stationary $ARMA(p,q)$ errors. *J. Econometrics* **47**, 333–357. MR1097742