

# Regression and modeling

Linear and multiple regression are among the most widely used statistical methods. This chapter considers empirical likelihood inferences for linear regression and other models with covariates, such as generalized linear models. The standard setting for linear regression has fixed predictors and a random response. Empirical likelihood was developed in Chapter 3.4 for smooth functions of means and for estimating equations, but assuming independent identically distributed data. Some new techniques are required to extend empirical likelihood to settings with fixed regressors.

Figure 4.1 shows a measure of breast cancer mortality versus population size for a set of counties in the southern U.S.A. A linear regression fits this data set well, though it is clear from the data, and obvious scientifically, that the variance of mortality increases with the population size. It is also reasonable to consider a regression through the origin for this set of data.

In simple linear regression, we observe pairs  $(X_i, Y_i)$ ,  $i = 1, \dots, n$  and it is thought that for a generic pair  $(X, Y)$ ,

$$E(Y | X = x) \doteq \beta_0 + \beta_1 x,$$

where the approximate inequality allows for some practically insignificant lack of fit. There are two widely used sampling models for linear regression. In one case  $(X_i, Y_i)$  are independent random vectors from a joint distribution  $F_{X,Y}$  on  $\mathbb{R}^2$ , while in the other,  $X_i = x_i$  are fixed and  $Y_i$  are then sampled independently from the conditional distributions  $F_{Y|X=x_i}$ . It is also a common practice to sample random pairs, but to analyze the data as if the  $X_i$  had been fixed at their observed values. These two sampling models extend to multiple regression in an obvious way.

This chapter first considers independent sampling of  $X, Y$  pairs because the results from Chapter 3 apply directly. Then sampling with fixed  $X_i$  is handled using a more general ELT. Then extensions are made to generalized linear models, nonlinear least squares, and the analysis of variance.

## 4.1 Random predictors

Suppose that  $X \in \mathbb{R}^p$  and  $Y \in \mathbb{R}$  are the generic predictor vector and response. Let  $(X_i, Y_i)$  be independent random observations from a common distribution. To make the notation simpler, suppose that  $X_i$  includes any necessary functions

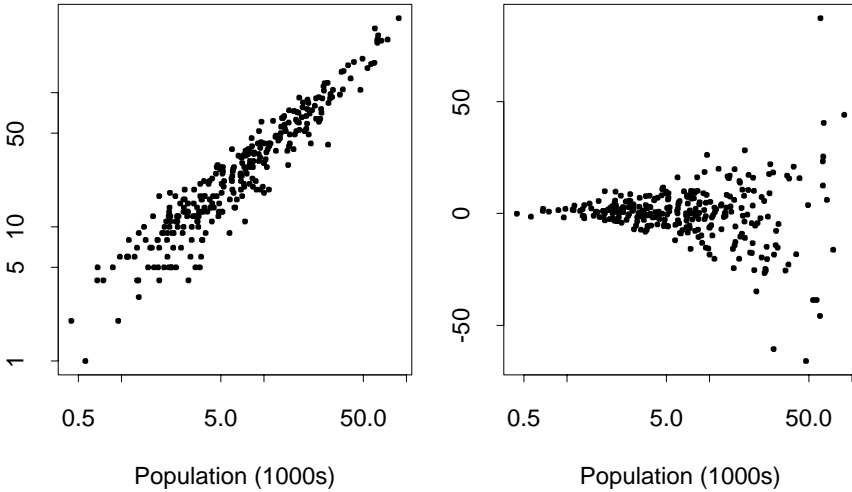


Figure 4.1 The left plot shows some cancer mortality counts (plus one) versus the population sizes for some counties in the southern U.S.A. The right plot shows residuals for these data from a linear regression of cancer on population. Source: Rice (1988).

of the available measurements. Thus  $X = (1, W)'$  for simple linear regression on  $W$ , and  $X = (1, U, W, U^2, W^2, UW)'$  for a quadratic response surface in  $U$  and  $W$ . We will suppose that  $E(X'X)$  has full rank  $p$ . It may be necessary to remove one or more redundant predictor variables to arrive at  $X$  with  $E(X'X)$  of full rank.

A linear model takes the form  $X'\beta$  for some vector  $\beta \in \mathbb{R}^p$ . The value  $\beta_{\text{LS}}$  that minimizes  $E((Y - X'\beta)^2)$  is

$$\beta_{\text{LS}} = E(X'X)^{-1} E(X'Y),$$

and the sample least squares estimate of  $\beta_{\text{LS}}$  is

$$\hat{\beta}_{\text{LS}} = \left( \frac{1}{n} \sum_{i=1}^n X_i'X_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n X_i'Y_i \right),$$

the NPMLE of  $\beta_{\text{LS}}$  in the sense of Chapter 2. We use  $\beta_{\text{LS}}$  instead of  $\beta_0$  to designate the population value of  $\beta$ , because in regression problems  $\beta_0$  has a firmly established use as an intercept coefficient.

Because  $\beta_{\text{LS}}$  is a smooth function of means, empirical likelihood inferences for it follow by the theory in Chapter 3.4, under mild moment conditions: that  $E(\|X\|^4) < \infty$ ,  $E(\|X\|^2 Y^2) < \infty$ , and  $E(X'X)$  is nonsingular. Invertibility of  $E(X'X)$  is needed in order to make  $\beta_{\text{LS}}$  a smooth function of means. There is

no need to assume that either  $X$  or  $Y$  has a normal distribution, or that  $\sigma^2(x) = \text{Var}(Y | X = x)$  is constant with respect to  $x$ .

There is also no need to assume that  $\mu(x) = E(Y | X = x)$  is of the form  $x'\beta$ . In this case the interpretation is that empirical likelihood confidence regions for the best value  $\beta_{\text{LS}}$  are properly calibrated, despite the lack of fit that might be inherent in that best value. A very large lack of fit could make the linear model irrelevant, and perhaps require the addition of some components to  $X$ . But a small lack of fit might be acceptable. In practice, some lack of fit is inevitable for many applications, and empirical likelihood tests are not sensitive to it.

The mean square prediction error  $E((Y - X'\beta_{\text{LS}})^2)$  is also a smooth function of means and so empirical likelihood inferences apply to it. This squared error can be written  $E(\sigma^2(X)) + E((\mu(X) - X'\beta_{\text{LS}})^2)$ , combining variance and lack of fit terms.

The regression model can also be approached through estimating equations. The definition of  $\beta_{\text{LS}}$  is equivalent to

$$E(X(Y - X'\beta_{\text{LS}})) = 0, \quad (4.1)$$

and the definition of  $\hat{\beta}_{\text{LS}}$  is equivalent to the normal equations

$$\frac{1}{n} \sum_{i=1}^n X_i(Y_i - X_i'\hat{\beta}_{\text{LS}}) = 0.$$

That is, the errors  $Y - X'\beta_{\text{LS}}$  are uncorrelated with  $X$  and the residuals  $Y_i - X_i'\hat{\beta}_{\text{LS}}$  are orthogonal to the sample  $X_i$ s. This formulation allows us to weaken the moment conditions for empirical likelihood regression inferences. The conditions  $E(\|X\|^4) < \infty$ , and  $E(\|X\|^2 Y^2) < \infty$  can be replaced by  $E(\|X\|^2 (Y - X'\beta_{\text{LS}})^2) < \infty$ . It is still necessary to have  $E(X'X)$  invertible so that  $\beta_{\text{LS}}$  is determined by (4.1).

Define the auxiliary variables  $Z_i = Z_i(\beta) = X_i(Y_i - X_i'\beta)$ . The empirical likelihood ratio function for  $\beta$  is defined by

$$\mathcal{R}(\beta) = \max \left\{ \prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i Z_i(\beta) = 0, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}.$$

For any vector  $\beta$  this empirical likelihood ratio may be computed using the algorithm for a vector mean, applied to the  $Z_i$  values. When a single component of  $\beta$  is of interest, then we maximize  $\mathcal{R}$  over the other components to obtain the profile empirical likelihood ratio function for the component of interest.

For the cancer data, we take  $X_i = (1, P_i)'$  where  $P_i$  is the population of the  $i$ 'th county, and  $Y_i = C_i$  where  $C_i$  is the number of cancer deaths in the  $i$ 'th county. Then  $\beta = (\beta_0, \beta_1)'$ . For these data  $\hat{\beta}_1 = 3.58$ , corresponding to a rate of 3.58 cancer deaths per 1000 population. Because deaths were counted over 20 years, the annualized rate is  $3.58/20 = 0.18$  per thousand. Also the intercept  $\hat{\beta}_0 = -0.53$  is quite close to zero, as we would expect.

Figure 4.2 shows the profile empirical log likelihood ratio functions for  $\beta_0$  and

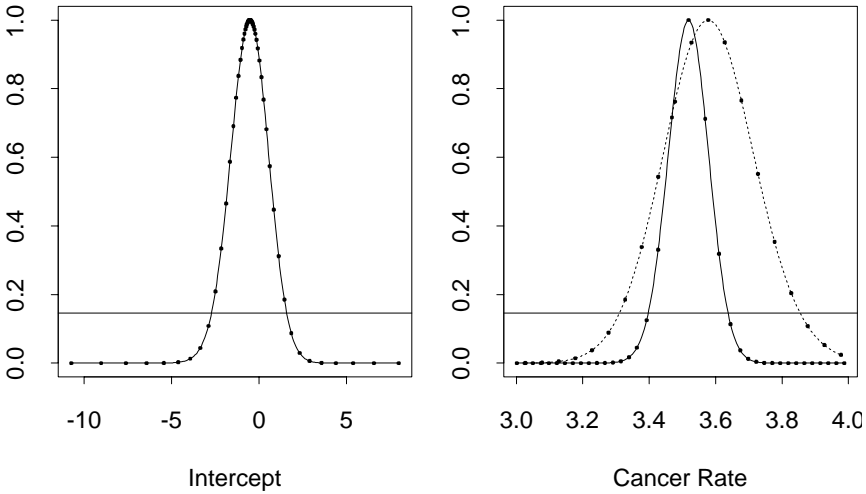


Figure 4.2 The left plot shows the empirical likelihood ratio function for the intercept in a simple linear regression model relating cancer mortality to county population for the data shown in Figure 4.1. A value of zero is reasonable for the intercept. The right plot shows the empirical likelihood ratio function for the slope. The solid curve is for regression through the origin, the dotted curve is for ordinary regression not through the origin.

$\beta_1$ . Horizontal reference lines mark asymptotic confidence thresholds. These data are clearly non-normal, and have nonconstant variance, but empirical likelihood inferences still have good coverage properties in this setting.

It is natural with data such as these to consider a regression through the origin. It is clear that a county with near zero population must have near zero cancer deaths. Furthermore, linearity appears to hold over the observed range of data, including some very small counties. To obtain a regression through the origin with slope  $\beta_1$  we insert  $\beta_0 = 0$  into the reweighted normal equations, obtaining

$$\sum_{i=1}^n w_i(C_i - P_i\beta_1) = 0, \quad \text{and} \quad (4.2)$$

$$\sum_{i=1}^n w_i P_i(C_i - P_i\beta_1) = 0. \quad (4.3)$$

This almost appears to be solving two equations in one unknown  $\beta_1$ , one to make the residuals have reweighted mean 0 and the other to make them uncorrelated with the population size. This is very different from what we would do with ordinary regression through the origin. In ordinary regression through the origin, the slope is estimated by  $\sum_{i=1}^n X_i Y_i / \sum_{i=1}^n X_i^2$ , for scalar  $X_i$ , corresponding to the second equation above. Then the resulting residuals do not sum to zero.

The profile empirical likelihood ratio  $\mathcal{R}(\beta_1)$  is found by maximizing  $\prod_{i=1}^n n w_i$  subject to  $w_i \geq 0$ ,  $\sum_{i=1}^n w_i = 1$ , and the estimating equations (4.2) and (4.3) above, for a fixed value of  $\beta_1$ . The maximizing value is taken to be  $\hat{\beta}_1$ . Under mild conditions  $-2 \log(\mathcal{R}(\beta_{LS,1})/\mathcal{R}(\hat{\beta}_1)) \rightarrow \chi_{(1)}^2$  in distribution as  $n \rightarrow \infty$ , when  $\beta_{LS,1}$  is the true value of  $\beta_1$  and the regression is truly through the origin.

Figure 4.2 also shows the profile empirical log likelihood ratio curve for the slope using regression through the origin. Because there is a correlation between the estimated intercept and slope, fixing the intercept at its known value makes a difference. This constraint shifts the MLE of the cancer rate down slightly to 3.52. It substantially narrows the likelihood ratio curve. As Figure 4.2 shows the confidence interval for  $\beta_{LS,1}$  using regression through the origin is less than half as wide as that without the constraint. Halving the length of a confidence interval usually requires a quadrupling of the sample size. Thus, one might interpret Figure 4.2 to mean that choosing regression through the origin more than quadruples the effective sample size  $n$ . We will revisit this issue, with another interpretation, in Chapter 4.5.

## 4.2 Nonrandom predictors

The usual model for linear regression has deterministic values  $X_i = x_i$ . Sometimes these values have been fixed by an experimental design. Sometimes they are fixed by conditioning. To fix  $X$  by conditioning, in a parametric model, proceed as follows. Factor the joint density or probability of  $X$  and  $Y$  as a product with one factor for the  $X$  distribution and another for the distribution of  $Y$  given  $X$ . Commonly the  $X$  distribution does not involve the regression parameters, and so all the information about them is in the conditional distribution of  $Y$  given  $X$ . Then it makes sense to use the conditional likelihood of  $Y_1, \dots, Y_n$  given  $X_1 = x_1, \dots, X_n = x_n$ , as in Chapter 3.8.

With  $X$  fixed, either by design or conditioning, the data are analyzed conditionally on the observed values of  $X$ . Now we suppose that  $E(Y_i) = \mu_i$  and that  $\text{Var}(Y_i) = \sigma_i^2$ .

With fixed regressors,  $Z_i(\beta) = x_i(Y_i - x_i'\beta)$  has mean  $x_i(\mu_i - x_i'\beta)$  and variance  $x_i x_i' \sigma_i^2$ . A linear model assumption  $\mu_i = x_i' \beta_{LS}$  makes the  $Z_i(\beta_{LS})$  independent with mean zero and nonconstant variance. They would still have nonconstant variance even if  $\sigma_i^2$  were constant, because of the factor  $x_i x_i'$ . Chapter 4.3 presents a triangular array ELT, for independent but not necessarily identically distributed random vectors, that applies to regression with fixed  $X_i$  when  $\mu_i = x_i' \beta$  holds for some  $\beta$ . As with random sampling of  $(X, Y)$  pairs, the main requirements are on moments.

Lack of fit is more serious in the fixed regressor setting, because it makes it problematic to define a true value  $\beta_{LS}$  upon which to draw inferences. A natural

default value to consider for  $\beta_{LS}$  is

$$\left( \sum_{i=1}^n x_i x_i' \right)^{-1} \left( \sum_{i=1}^n x_i \mu_i \right),$$

corresponding to random regressors sampled from the empirical distribution on  $X$ . Empirical likelihood inferences in this setting tend to be conservative. Asymptotically they cover the true value more often than the nominal level indicates.

**Theorem 3.5** shows that imposing the constraint  $\sum w_i U_i = E(U)$  for sample values  $U_i$  has asymptotically the same effect as conditioning on the observed value of  $(1/n) \sum_{i=1}^n U_i$ . So this suggests that we might be able to capture the effects of conditioning on  $x_i$  by employing the constraints

$$\sum_{i=1}^n w_i x_i = \frac{1}{n} \sum_{i=1}^n x_i$$

and

$$\sum_{i=1}^n w_i x_i x_i' = \frac{1}{n} \sum_{i=1}^n x_i x_i'$$

while computing  $\mathcal{R}(\beta)$ . Then our analysis will, in the sense of **Theorem 3.5**, be conditional on the first and second moments of  $x$ .

To match even more features of the observed  $x$  values, let  $Q(x)$  be an  $s$ -dimensional function of  $x$ , and consider imposing the constraint  $\sum_{i=1}^n w_i Q(x_i) = (1/n) \sum_{i=1}^n Q(x_i) = \bar{Q}$  analogous to conditioning on the observed value of  $\bar{Q}$ . These  $s$  constraints cannot widen the confidence region for  $\beta$ . They would narrow it asymptotically if there were any correlation between  $Q(X_i)$  and  $Z_i(\beta)$ .

When  $\beta_{LS}$  is the true regression vector,

$$E(Z_i(\beta_{LS})Q(X_i)) = E(E(Z_i(\beta_{LS}) | X_i)Q(X_i)) = 0.$$

There appears to be no advantage to conditioning on any finite set of moments of  $X$  in the limit as  $n \rightarrow \infty$ . For a scalar function of  $\beta_{LS}$ , the ratio of constrained to unconstrained confidence interval lengths approaches 1.0 as  $n \rightarrow \infty$ . Perhaps there are benefits to imposing constraints on the reweighted  $x_i$ , but if so they tend to disappear as  $n \rightarrow \infty$ .

There are two common reasons for fixing regressors by conditioning. The first is that the conditional variance of the least squares estimator  $\hat{\beta}_{LS}$  is simpler than the unconditional one, especially when the errors are assumed to have constant variance.

The second, and more subtle, reason is based on the statistical idea of ancillarity. Suppose that observing the  $X_i$  does not give us any information about the value of  $\beta_{LS}$  but does give us information about the variance of  $\hat{\beta}_{LS}$ . For instance, with constant error variance,  $\text{Var}(\hat{\beta}_{LS}) = (\sum_{i=1}^n x_i x_i')^{-1} \sigma^2$ . A statistic like  $\sum_{i=1}^n x_i x_i'$ , which tells us nothing about  $\beta_{LS}$  but something about  $\text{Var}(\hat{\beta}_{LS})$ ,

is called ancillary. The point of conditioning on an ancillary statistic is to use the known variance of our estimate rather than considering other  $x_i$  and corresponding other  $\text{Var}(\hat{\beta}_{\text{LS}})$ 's that we might have had instead. In particular, we should expect confidence regions for  $\beta_{\text{LS}}$  to be large when  $\sum_{i=1}^n x_i x_i'$  is small and vice versa. This indeed holds for empirical likelihood, and most if not all other widely used confidence interval methods for regression.

### 4.3 Triangular array ELT

In applications such as linear regression with nonrandom predictors, and kernel smoothing (Chapter 5), empirical likelihood is applied to the mean of random variables that are not necessarily independent and identically distributed. Instead a triangular array structure  $Z_{in} \in \mathbb{R}^p$ , for  $i = 1, \dots, n$  is appropriate. The vectors can be arranged into a triangular array,

$$\begin{array}{cccc} Z_{11} & & & \\ Z_{12} & Z_{22} & & \\ Z_{13} & Z_{23} & Z_{33} & \\ \vdots & \vdots & \vdots & \ddots \\ Z_{1n} & Z_{2n} & Z_{3n} & \cdots & Z_{nn}. \end{array}$$

For each  $n$ , we will assume that  $Z_{1n}, \dots, Z_{nn}$  are independent, but not necessarily that they are identically distributed. For regression with fixed regressors,  $Z_{in}(\beta) = x_i(Y_i - x_i'\beta)$ . In regression  $Z_{in} = Z_{im}$  for  $i \leq n < m$ , though this does not hold in other settings. We assume that  $E(Z_{in})$  is the same for  $i = 1, \dots, n$ . In regression, this common value is 0 for all  $n$ , but in other settings the common mean can depend on  $n$ . The variances of the  $Z_{in}$  have to be of roughly the same order of magnitude for a central limit theorem to hold.

Introduce  $V_{in} = \text{Var}(Z_{in})$  and  $V_n = (1/n) \sum_{i=1}^n V_{in}$ , and for a real symmetric matrix  $A$ , let  $\text{maxeig}(A)$  and  $\text{mineig}(A)$  denote the largest and smallest eigenvalues of  $A$ , respectively.

**Theorem 4.1 (Triangular array ELT)** *Let  $Z_{in} \in \mathbb{R}^p$  for  $1 \leq i \leq n$  and  $n \geq n_{\min}$  be a triangular array of random vectors. Suppose that for each  $n$ , that  $Z_{1n}, \dots, Z_{nn}$  are independent and have common mean  $\mu_n$ . Let  $\mathcal{H}_n$  denote the convex hull of  $Z_{1n}, \dots, Z_{nn}$ , and put  $\sigma_{1n} = \text{maxeig}(V_n)$ , and  $\sigma_{pn} = \text{mineig}(V_n)$ . Assume that as  $n \rightarrow \infty$*

$$\Pr(\mu_n \in \mathcal{H}_n) \rightarrow 1 \tag{4.4}$$

and

$$\frac{1}{n^2} \sum_{i=1}^n E \left( \|Z_{in} - \mu_n\|^4 \sigma_{1n}^{-2} \right) \rightarrow 0 \tag{4.5}$$

and that for some  $c > 0$  and all  $n \geq n_{\min}$ ,

$$\frac{\sigma_{pn}}{\sigma_{1n}} \geq c. \quad (4.6)$$

Then  $-2 \log \mathcal{R}(\mu_n) \rightarrow \chi_{(p)}^2$  in distribution as  $n \rightarrow \infty$ , where

$$\mathcal{R}(\mu_n) = \max \left\{ \prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i (Z_{in} - \mu_n) = 0, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}.$$

*Proof.* This theorem is proved in Chapter 11.3.  $\square$

For regression problems with fixed predictors  $x_i$  and independent responses  $Y_i$ , the vectors  $Z_{in}(\beta_0) = x_i(Y_i - x_i'\beta_0)$  are independent. They all have mean 0 if  $\beta_0$  is the true common regression vector. That is, if  $E(Y \mid X = x_i) = x_i'\beta_0$ , for  $i = 1, \dots, n$ .

The asymptotic results do not depend on which  $n_{\min}$  is used. For regression problems  $n_{\min} \geq p$ , for otherwise  $\sigma_{pn} = 0$ . In applications there are usually many more data points than regression coefficients, so the value  $n_{\min}$  has only a minor role.

The convex hull condition (4.4), which was easily satisfied for the mean of IID random vectors, must be investigated separately for each use of the triangular array ELT. If all the errors  $e_i$  are positive then it is certain that  $0 \notin \mathcal{H}_n$ . Similarly, for simple linear regression with  $x_i = (1, t_i)'$ , if the regression  $\beta_0 + \beta_1 t$  is increasing, with  $\beta_1 > 0$ , but the sample data are decreasing, with  $Y_i < Y_j$  whenever  $t_i > t_j$ , then  $0 \notin \mathcal{H}_n$ . We cannot reweight a decreasing sample to get an increasing regression line.

Under normal circumstances, we expect however that (4.4) will be satisfied rapidly unless  $p$  is large or  $n$  is small. Let  $e_i = Y_i - x_i'\beta_0$  be the error for observation  $i$ , so that  $Z_{in} = x_i e_i$ . We require that some vector of weights  $w_i \geq 0$  exists with  $\sum_{i=1}^n w_i = 1$  and  $\sum_{i=1}^n w_i x_i e_i = 0$ . Here is a simple sufficient condition.

**Lemma 4.1** *Let  $\mathcal{H}_n^+$  be the convex hull of the set  $\{x_i \mid e_i > 0, 1 \leq i \leq n\}$  and let  $\mathcal{H}_n^-$  be the convex hull of the set  $\{x_i \mid e_i < 0, 1 \leq i \leq n\}$ . If  $\mathcal{H}_n^+ \cap \mathcal{H}_n^- \neq \emptyset$ , then  $0 \in \mathcal{H}_n$ .*

*Proof.* If  $x \in \mathcal{H}_n^+ \cap \mathcal{H}_n^-$ , then we may write  $x = \sum_{i=1}^n w_i^+ x_i = \sum_{i=1}^n w_i^- x_i$  with all  $w_i^\pm \geq 0$ ,  $\sum_{i=1}^n w_i^\pm = 1$ ,  $w_i^+ = 0$  if  $e_i \leq 0$ , and  $w_i^- = 0$  if  $e_i \geq 0$ . Then  $\sum_{i=1}^n w_i^\pm Z_{in} = x \tilde{e}_\pm$  where  $\tilde{e}_\pm = \sum_{i=1}^n w_i^\pm e_i$ . We have  $\tilde{e}_- < 0 < \tilde{e}_+$ , and by taking  $w_i = (w_i^+ |\tilde{e}_-| + w_i^- |\tilde{e}_+|) / (|\tilde{e}_+| + |\tilde{e}_-|)$ , we find

$$\sum_{i=1}^n w_i Z_{in} = \frac{|\tilde{e}_-| x \tilde{e}_+ + |\tilde{e}_+| x \tilde{e}_-}{|\tilde{e}_+| + |\tilde{e}_-|} = 0,$$

and so  $0 \in \mathcal{H}_n$ .  $\square$

For simple linear regression with  $Y_i = \beta_0 + \beta_1 t_i + e_i$  and  $x_i = (1, t_i)'$ , the convex hull condition simplifies further. Let  $I_n^+$  be the interval from the smallest  $t_i$  with  $e_i > 0$  to the largest  $t_i$  with  $e_i > 0$ , and let  $I_n^-$  be similarly defined using  $t_i$  with  $e_i < 0$ . If these intervals overlap then  $0 \in \mathcal{H}_n$ . We simply require a triple  $t_i < t_j < t_k$  where the sign of  $e_j$  differs from those of  $e_i$  and  $e_k$ . If on the other hand there is some value  $t$  such that  $e_i > 0$  whenever  $t_i > t$  and  $e_i < 0$  whenever  $t_i < t$ , then  $0$  might not be in  $\mathcal{H}_n$ .

The vectors  $Z_{in}$  have variance  $V_{in} = x_i x_i' \sigma_i^2$ , where  $\sigma_i^2 = \text{Var}(Y_i)$ , so

$$V_n = \frac{1}{n} \sum_{i=1}^n x_i x_i' \sigma_i^2.$$

Under mild conditions on  $x_i$  and  $\sigma_i^2$ , both  $\sigma_{1n} = \text{maxeig}(V_n)$ , and  $\sigma_{pn} = \text{mineig}(V_n)$  have finite nonzero limits,  $\sigma_{1\infty}$  and  $\sigma_{p\infty}$ , respectively, as  $n \rightarrow \infty$ .

For regression, condition (4.5) becomes

$$\frac{1}{n^2} \sum_{i=1}^n \|x_i\|^4 E(e_i^4) \sigma_{1n}^{-2} \rightarrow 0.$$

If  $\sigma_{1n}$  tends to a finite nonzero limit, then a sufficient condition for (4.5) is that  $\max_{1 \leq i \leq n} E(e_i^4)$  and  $\max_{1 \leq i \leq n} \|x_i\|^4$  both have a finite upper bound holding for all  $n$ . Still weaker conditions are sufficient. These quantities are allowed to diverge slowly to infinity, and even the average of  $\|x_i\|^4 E(e_i^4)$  can diverge to infinity as long as it remains  $o(n)$ .

For condition (4.6), the constant  $c$  can be taken to be slightly smaller than  $\sigma_{p\infty}/\sigma_{1\infty}$  when these exist. To violate condition (4.6) would require either unbounded ratios of observation variances  $\sigma_i^2$  or unbounded ratios of extreme eigenvalues of  $n^{-1} \sum_{i=1}^n x_i x_i'$ .

#### 4.4 Analysis of variance

The one way analysis of variance (ANOVA) is widely used to compare the means of different populations. Suppose that we observe independent random variables  $Y_{ij} \in \mathbb{R}$ , for  $i = 1, \dots, k$  and  $j = 1, \dots, n_i$ . The groups  $i$  are considered to have possibly different means, and usually an identical variance. Then the null hypothesis of identical means is rejected when the ratio

$$\frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}{\frac{1}{N-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2}$$

exceeds the  $1 - \alpha$  quantile  $F_{k-1, N-k}^{1-\alpha}$  of the  $F_{k-1, N-k}$  distribution. Here  $N = \sum_{i=1}^k n_i$ , and  $\bar{Y}_{i\bullet} = (1/n_i) \sum_{j=1}^{n_i} Y_{ij}$ , and  $\bar{Y}_{\bullet\bullet} = (1/N) \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$ . The  $F$  distribution holds if the  $Y_{ij}$  are normally distributed, and it holds asymptotically for non-normal data. The assumption of a common variance is critical for asymptotic validity, unless the  $n_i$  are equal or nearly so.

For an empirical likelihood approach to ANOVA, we suppose that  $Y_{ij} \in \mathbb{R}^d$  are independent and have the distribution  $F_{i0}$ , for  $d \geq 1$ . We do not need to distinguish ANOVA ( $d = 1$ ) from multivariate ANOVA (MANOVA) with  $d > 1$ . A natural approach to empirical likelihood for this setting is to define the likelihood function

$$L_k(F_1, \dots, F_k) = \prod_{i=1}^k \prod_{j=1}^{n_i} v_{ij},$$

where  $v_{ij} = F_i(\{Y_{ij}\})$ . The empirical likelihood ratio function is then

$$R_k(F_1, \dots, F_k) = \prod_{i=1}^k \prod_{j=1}^{n_i} n_i v_{ij}.$$

An alternative formulation is to encode the data as  $N$  pairs  $(I, Y)$  where  $I \in \{1, \dots, k\}$  and  $Y \in \mathbb{R}^d$ . The observation  $Y_{ij}$  is represented by a pair with  $I = i$  and  $Y = Y_{ij}$ . Let  $F$  be a distribution on  $(I, Y)$  pairs. The data are not an IID sample from any such distribution  $F_0$ , in the usual setting where  $n_i$  are nonrandom. Instead the variable  $I$  behaves more like a nonrandom categorical predictor. Define the likelihood

$$L(F) = \prod_{i=1}^k \prod_{j=1}^{n_i} w_{ij},$$

where  $w_{ij} = F(\{(i, Y_{ij})\}) \geq 0$  and  $\sum_{i=1}^k \sum_{j=1}^{n_i} w_{ij} = 1$ .

The weights  $w_{ij}$  can be factored into  $w_{j|i} w_{i\bullet}$  where  $w_{i\bullet} = \sum_{j=1}^{n_i} w_{ij}$  and  $w_{j|i} = w_{ij}/w_{i\bullet}$ . The  $w_{i\bullet}$  factor describes the probability attached by  $F$  to group  $I = i$ , while the factor  $w_{j|i}$  describes the distribution of  $Y_{Ij}$  given that  $I = i$ . In ANOVA problems, we are usually interested in  $F$  only through  $w_{j|i}$ . This is necessarily true when the  $n_i$  have been fixed by the experimental design. The empirical likelihood ratio function on data pairs may be written

$$\begin{aligned} R(F) &= \prod_{i=1}^k \prod_{j=1}^{n_i} N w_{i\bullet} w_{j|i} \\ &= \left( \prod_{i=1}^k \left( \frac{N w_{i\bullet}}{n_i} \right)^{n_i} \right) \left( \prod_{i=1}^k \prod_{j=1}^{n_i} n_i w_{j|i} \right). \end{aligned}$$

If we maximize  $R(F)$  subject to constraints that only involve  $w_{j|i}$ , then the result will have  $w_{i\bullet} = n_i/N$  and so

$$R(F) = R_k(F_1, \dots, F_k)$$

where  $F_i(\{Y_{ij}\}) = w_{j|i}$ . Maximizing the likelihood ratio for such constraints automatically keeps the group weights  $w_{i\bullet}$  proportional to the actual sample sizes  $n_i$ .

Empirical likelihood confidence regions and tests for the one way ANOVA will

be identical whether they are constructed through  $R(F)$  or through the more natural  $R_k(F_1, \dots, F_k)$ , as long as the statistic  $T(F)$  depends only on the conditional distributions of  $Y_{Ij}$  given  $I = i$  and not on the marginal distribution of the group variable  $I$ . The triangular array ELT is then available to justify the method based on sampling  $(I, Y)$  pairs.

Suppose that  $\mu_{i0} = \int y dF_{i0}(y) \in \mathbb{R}^d$  and define

$$\mathcal{R}(\mu_1, \dots, \mu_k) = \max \left\{ \prod_{i=1}^k \prod_{j=1}^{n_i} N w_{ij} \mid w_{ij} \geq 0, \sum_{i=1}^k \sum_{j=1}^{n_i} w_{ij} = 1 \right. \\ \left. \sum_{j=1}^{n_i} w_{ij} (Y_{ij} - \mu_i) = 0, j = 1, \dots, k, \right\}.$$

To apply the triangular array ELT, define the auxiliary variables  $Z_{ijN} \in \mathbb{R}^D$ , where  $D = kd$ . Taking  $Y_{ij}$  to be column vectors, we write

$$Z_{ijN} = (0, \dots, 0, Y'_{ij} - \mu'_i, 0, \dots, 0)',$$

where  $Y'_{ij} - \mu'_i$  is preceded by  $(i-1)d$  zeros and followed by  $(k-i)d$  zeros. We could rewrite  $Z_{ijN}$  as  $Z_{lN}$  for  $1 \leq l \leq N$  in order to make the notation more closely match that of the triangular array ELT, but this is not necessary. The key quantity in applying that theorem is the matrix  $V_N$  given by

$$V_N = \frac{1}{N} \begin{pmatrix} n_1 \text{Var}(Y_{11}) & 0 & \cdots & 0 \\ 0 & n_2 \text{Var}(Y_{21}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_k \text{Var}(Y_{k1}) \end{pmatrix}.$$

If each  $\text{Var}(Y_{i1})$  is finite and nonsingular, then the condition on the eigenvalues of  $V_N$  is satisfied so long as the sample sizes grow subject to

$$\lim_{N \rightarrow \infty} \frac{\min_i n_i}{\max_i n_i} > 0. \quad (4.7)$$

The convex hull condition for  $Z_{ijN}$  becomes  $k$  convex hull conditions: for each  $i = 1, \dots, k$ , the convex hull of  $Y_{ij}$  needs to contain  $\mu_{i0}$ . Thus under very mild conditions,

$$-2 \log \mathcal{R}(\mu_{10}, \dots, \mu_{k0}) \rightarrow \chi^2_{(D)}$$

in distribution as  $N \rightarrow \infty$ .

Returning to ANOVA, suppose that  $d = 1$ . Then

$$-2 \log \mathcal{R}(\mu_{10}, \dots, \mu_{k0}) \rightarrow \chi^2_{(k)}$$

in distribution as  $N \rightarrow \infty$ . The most common hypothesis is that  $\mu_{i0} = \mu_0$  all take the same (unknown) value. This hypothesis corresponds to  $k-1$  constraints on

the mean of  $Z_{ijN}$  instead of  $k$  constraints and so

$$-2 \max_{\mu} \log \mathcal{R}(\mu, \dots, \mu) \rightarrow \chi_{(k-1)}^2$$

in distribution as  $N \rightarrow \infty$  when  $\mu_{10} = \dots = \mu_{k0}$ . The observations do not need to be normally distributed, or to have a common variance, and the sample sizes need not be equal. Each group needs a finite nonzero variance.

For  $d > 1$ , empirical likelihood produces MANOVA tests for equality of  $\mu_i$  requiring very weak assumptions and having an asymptotic  $\chi_{(d(k-1))}^2$  calibration. The formulation above can be further generalized to a setting where the data have possibly different dimensions  $d_i$  in each population. An asymptotic  $\chi_D^2$  result holds for  $-2 \log \mathcal{R}(\mu_1, \dots, \mu_k)$  under conditions including a bound on the ratio of eigenvalues of the matrix  $V_N$ . Now the matrix  $V_N$  is  $D$  by  $D$  where  $D = \sum_{i=1}^k d_i$ . When  $d_1 \neq d_2$ , it is not natural to compare  $\mu_{10}$  and  $\mu_{20}$ , though comparisons of functions of  $\mu_1, \dots, \mu_k$  may be of interest.

More general multi-sample statistics are considered in Chapter 11.4. For example,  $\Pr(Y_{1j} > Y_{2j})$  is covered there, but not by the ANOVA formulation described above.

#### 4.5 Variance modeling

Least squares inferences can be inefficient when the response  $Y_i$  has nonconstant variance  $\sigma_i^2$ , for fixed regressors, or when  $\sigma^2(x) = \text{Var}(Y | X = x)$  is nonconstant in  $x$ , for random regressors. Greater accuracy can be obtained by weighting observation  $i$  in inverse proportion to the variance of  $Y_i$ . In some cases the inefficiency of unweighted least squares is mild and may be tolerated. In others, the inefficiency may be large enough that we seek to put more weight on the less variable observations. This can be done by introducing a model for the variance of  $Y$ . Sometimes we may introduce such a model because the variance is interesting in its own right.

Suppose that we observe  $(X, Z, Y)$  triples, where  $X$  and  $Z$  are thought to be related to the mean and variance of  $Y$ , respectively. It may be that  $Z$  is  $X$ , a subvector of  $X$ , or a transformation of  $X$ . Perhaps  $(X_i, Z_i, Y_i)$  are IID vectors, or alternatively  $x_i$  and  $z_i$  are fixed. Because the variance of  $Y$  cannot be negative it is natural to model the logarithm of the variance of  $Y$  using  $Z$ . The model  $Y | (X, Z) \sim N(X'\beta, \exp(2Z'\gamma))$ , leads to the estimating equations

$$0 = \frac{1}{n} \sum_{i=1}^n \exp(-2z_i'\gamma) x_i (Y_i - x_i'\beta) \tag{4.8}$$

$$0 = \frac{1}{n} \sum_{i=1}^n z_i \left( 1 - \exp(-2z_i'\gamma) (Y_i - x_i'\beta)^2 \right). \tag{4.9}$$

For the breast cancer data, take  $x_i = (1, P_i)'$ , and  $z_i = (1, \log P_i)'$ , where  $P_i$  is the population of the  $i$ th county, and take  $Y_i = C_i$ , the number of cancer deaths

in that county. The model for the expected value of  $C_i$  is now  $\beta_0 + \beta_1 P_i$ , but we will run the regression through the origin, by constraining  $\beta_0 = 0$ . The model for the conditional standard deviation of  $C_i$  is a power model  $\exp(\gamma_0 + \gamma_1 \log(P_i)) = \exp(\gamma_0) P_i^{\gamma_1}$ . A normal distribution clearly cannot hold for discrete data, but the parameters  $\beta_1$ ,  $\gamma_0$ , and  $\gamma_1$  retain their interpretations as parameters of the mean and variance of  $C$  given  $P$ . Specifically,  $E(C|P) = \beta_1 P$ , and  $\text{Var}(Y|P) = \exp(2\gamma_0) P^{2\gamma_1}$ .

The simplest parametric model one might believe to hold for cancer incidence is Poisson with intensity proportional to the population size. This model would give a regression through the origin. It would also give  $\gamma_1 = 1/2$  and  $\gamma_0 = \log(\beta_1)/2$ , because for a Poisson model the variance equals the mean. It is often found that a Poisson model fails to fit epidemiological data. A Poisson model may be derived by assuming that different people get cancer independently of each other, and with the same small probability. Commonly there is overdispersion, wherein the variance of observations with different means increases faster than does the mean. This can happen with clustering (as for families within counties) or because there is variation from county to county in factors such as industrial exposure, age, or smoking. A commonly used model for data overdispersed relative to the Poisson is the Gamma distribution. For Gamma distributions,  $\gamma_1 = 1$ .

The MLE's for the cancer data are  $\hat{\beta}_1 = 3.57$ ,  $\hat{\gamma}_0 = 0.602$  and  $\hat{\gamma}_1 = 0.731$ . The value for  $\gamma_1$  describes overdispersion relative to a Poisson model, but less overdispersion than would hold in a Gamma model.

The left panel of [Figure 4.3](#) shows the profile empirical likelihood ratio function for  $\beta_1$ . For comparison purposes, the curve for  $\beta_1$  from a regression through the origin without variance modeling is also shown. The two curves have roughly the same width. The approximate four-fold efficiency gain from regression through the origin can also be obtained by variance modeling which puts more weight on observations from small counties. Also, using both techniques is not much more accurate than using just one of them.

The right panel of [Figure 4.3](#) plots the profile empirical likelihood function for  $\gamma_1$  using both regression through the origin and unconstrained regression. The constraint narrows the empirical likelihood function peak slightly. In particular, constraining the regression to go through the origin raises the lower confidence limit for  $\gamma_1$  somewhat. By either curve, we can infer that both the overdispersion relative to the Poisson model and the underdispersion relative to the Gamma model are statistically significant. The value 0.75 is near the center of the confidence interval for  $\gamma_1$ , and this corresponds to weighting the observations proportionally to  $P^{1.5}$  instead of  $P$  or  $P^2$  as in Poisson and Gamma models, respectively.

## 4.6 Nonlinear least squares

In some applications there is a specific functional form  $f(x, \theta)$  that is known or suspected to give  $E(Y|X = x)$ . Then a reasonable way to estimate  $\theta$  is to

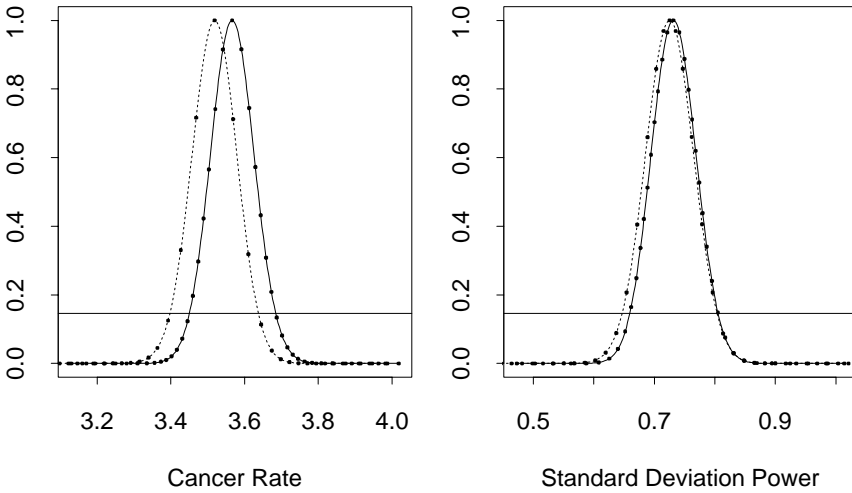


Figure 4.3 The left plot shows the empirical likelihood function for the intercept in a linear regression through the origin, relating cancer mortality to county population for the data shown in Figure 4.1. The solid curve is from a heteroscedastic regression model, the dotted curve is from a regression model using constant variance. The right plot shows the empirical likelihood function for the exponent in a power law relating the standard deviation of the cancer mortality level to the population size. Values 1/2 and 1, corresponding to Poisson and Gamma models, respectively, do not fit the data. The solid curve is for regression through the origin, the dotted curve is for regression not constrained to pass through the origin.

minimize

$$S(\theta) = \sum_{i=1}^n (Y_i - f(x_i, \theta))^2 \tag{4.10}$$

with respect to  $\theta$ . Minimizing equation (4.10) gives the maximum likelihood estimate of  $\theta$  under a model with independent  $Y_i \sim N(f(x_i, \theta), \sigma^2)$ . We suppose that  $f(x, \theta)$  is not linear in  $\theta$ , for otherwise linear regression methods could be used.

The least squares estimate is still valuable, even if  $Y_i$  are not normally distributed or do not have equal variance. For random  $X_i$ , the population quantity estimated is that value of  $\theta$  minimizing  $E((Y - f(X, \theta))^2)$ . In particular, if  $E(Y|X = x)$  were really of the form  $f(x, \theta_0)$ , then  $\theta_0$  minimizes  $E((Y - f(X, \theta))^2)$ .

Finding the nonlinear least squares estimate  $\hat{\theta}$  can be a challenge. Success can depend on using good starting values, and it may be necessary to rescale the data or parameter values to avoid loss of numerical accuracy. The same estimate  $\hat{\theta}$  is

used for empirical likelihood and parametric inferences. The methods differ when it comes to constructing confidence sets.

Normal theory confidence regions for  $\theta$  may be obtained either by linearizing the model or by profiling the log likelihood function. Linearization inferences begin with the approximation

$$f(x_i, \theta) \doteq f(x_i, \theta_0) + (\theta - \theta_0)'g(x_i, \theta_0),$$

where

$$g(x_i, \theta) = \frac{\partial}{\partial \theta} f(x_i, \theta).$$

The vector  $\theta - \theta_0$  may be approximated by a linear regression with an  $n$  by  $p$  predictor matrix  $J$  and a response vector  $Z$  given by

$$J = J(\theta_0) = \begin{pmatrix} g(x_1, \theta_0)' \\ g(x_2, \theta_0)' \\ \vdots \\ g(x_n, \theta_0)' \end{pmatrix} \text{ and } Z = Z(\theta_0) = \begin{pmatrix} Y_1 - f(x_1, \theta_0) \\ Y_2 - f(x_2, \theta_0) \\ \vdots \\ Y_n - f(x_n, \theta_0) \end{pmatrix}.$$

The value  $\theta_0$  is unknown. The Gauss-Newton iteration estimates  $\theta_0$  by iterated least squares, replacing  $\hat{\theta}$  by  $\hat{\theta} + (\hat{J}'\hat{J})^{-1}\hat{J}'\hat{Z}$ , where  $\hat{J} = J(\hat{\theta})$  and  $\hat{Z} = Z(\hat{\theta})$ .

Under standard assumptions, the asymptotic distribution of  $n^{1/2}(\hat{\theta} - \theta_0)$  is  $N(0, \sigma^2(J'J)^{-1})$ . This matches our expectations under a linear regression model on  $J$ , and justifies ellipsoidal confidence regions for  $\theta$  of the form

$$\left\{ \theta \mid (\theta - \hat{\theta})' \hat{J}' \hat{J} (\theta - \hat{\theta}) \leq s^2 p F_{p, n-p}^{1-\alpha} \right\} \quad (4.11)$$

where  $s^2 = S(\hat{\theta})/(n - p)$ .

Confidence regions for  $\theta_0$  formed by thresholding the normal theory likelihood reduce to thresholding the function  $S(\theta)$ . The asymptotic distribution of  $(S(\theta_0) - S(\hat{\theta}))/ps^2$  is  $F_{p, n-p}$  justifying confidence regions of the form

$$\left\{ \theta \mid S(\theta) \leq S(\hat{\theta}) \left[ 1 + \frac{p}{n - p} F_{p, n-p}^{1-\alpha} \right] \right\} \quad (4.12)$$

It can be very difficult to get accurate inferences in nonlinear least squares problems. The source of this difficulty is curvature of the vector

$$f_n(\theta) = (f(x_1, \theta), \dots, f(x_n, \theta))'$$

expressed as a function of  $\theta$ . This mapping from  $\mathbb{R}^p$  to  $\mathbb{R}^n$  has two kinds of curvature: intrinsic curvature and parameter effects curvature. Intrinsic curvature arises because the  $p$ -dimensional surface  $\{f_n(\theta) \mid \theta \in \mathbb{R}^p\}$  in  $n$ -dimensional space is not flat but curved. Parameter effects curvature arises because of the labeling through  $\theta$  of the points in this surface. In a linear mapping, changing  $\theta$  to  $\theta + \Delta$  produces the same change in  $f_n$  for any value of  $\theta$ . In a nonlinear mapping

$f_n(\theta + \Delta) - f_n(\theta)$  can depend strongly on  $\theta$  for fixed  $\Delta$ . Linearization inferences implicitly approximate  $f_n(\theta + \Delta) - f_n(\theta)$  at every  $\theta$  by  $f_n(\hat{\theta} + \Delta) - f_n(\hat{\theta})$ , while methods based directly on the sum of squares do not make this approximation.

If we reparameterize, replacing  $\theta$  by  $\tau(\theta)$  the intrinsic curvature of the surface is unchanged, but the parameter effects curvature will usually have changed. It is common for nonlinear models to employ the exponential function, or even to have the exponential in an exponent. These models can have very high parameter effects curvature.

It has been found empirically that confidence regions for  $\theta$  and functions of  $\theta$  found by thresholding (profiling) the sum of squares are usually well calibrated but that confidence regions based on linearization can be very badly calibrated. The usual explanation is that intrinsic curvatures are typically small compared to parameter effects curvatures.

Another explanation for the success of methods based on the sum of squares runs as follows. For normally distributed  $Y_i$  with known and constant  $\sigma^2$ , the set  $\{\theta \mid S(\theta) \leq \sigma^2 \chi_{(n)}^{2, 1-\alpha}\}$  has exactly  $1 - \alpha$  coverage probability, regardless of the form of  $f$ . Coverage error can enter when  $\sigma^2$  is estimated. In a normal linear model, the average squared residual is  $S(\hat{\theta})/n = (n - p)s^2/n \sim \sigma^2 \chi_{(n-p)}^2/n$ . This average squared residual tends to be smaller than  $\sigma^2$ , but in a way that is well understood and easily corrected. For a pathological nonlinear model, in which the  $p$ -dimensional hyper-surface  $\{f_n(\theta) \mid \theta \in \mathbb{R}^p\} \subset \mathbb{R}^n$  nearly fills the  $n$ -dimensional space  $S(\hat{\theta})$  can be far smaller than  $\sigma^2$ , and no practical correction is available. But for reasonable models, the correction of  $S(\hat{\theta})$  implicit in (4.12) does not go far wrong.

Empirical likelihood inferences for nonlinear least squares are based on

$$\theta(w_1, \dots, w_n) = \arg \min_{\theta} \sum_{i=1}^n w_i (Y_i - f(x_i, \theta))^2.$$

If the sum of squares takes a unique minimum at a point where its gradient with respect to  $\theta$  vanishes, then the estimating equations

$$\sum_{i=1}^n w_i (Y_i - f(x_i, \theta)) g(x_i, \theta) = 0$$

serve to define  $\theta$  for given weights  $w_i$ . Because empirical likelihood is parameterization invariant, it is only affected by parameter effects curvature, in the same way that a parametric likelihood is. Empirical likelihood inferences have an advantage in not requiring constant error variance.

Figure 4.4 shows data measuring calcium uptake  $Y$  versus time  $X$ . A reasonable model for these data is

$$E(Y_i | X_i = x_i) = \theta_1 (1 - e^{-\theta_2 x_i}). \quad (4.13)$$

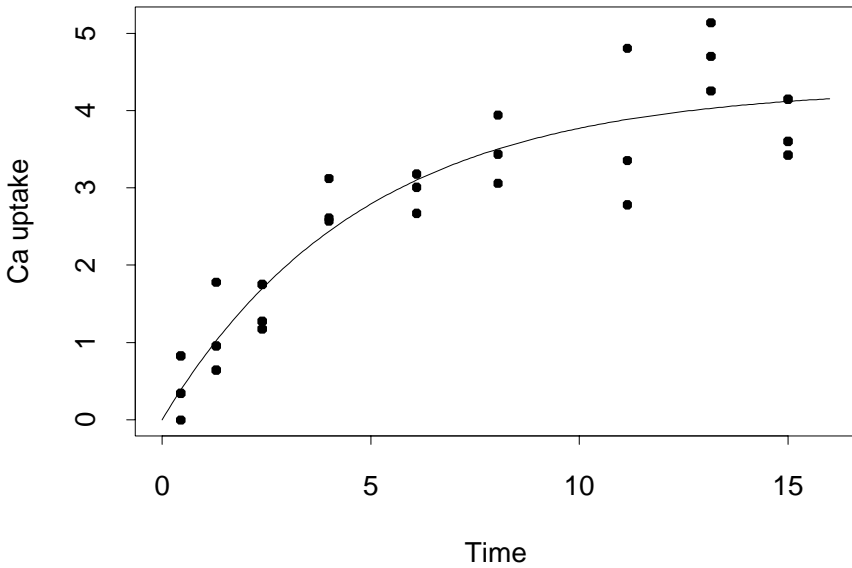


Figure 4.4 Shown are the calcium uptake data (Rawlings 1988) and a model  $E(Y|X = x) = \theta_1(1 - \exp(-\theta_2x))$  fit to the data by nonlinear least squares.

The least squares estimates are  $\hat{\theta}_1 = 4.309$  and  $\hat{\theta}_2 = 0.208$ . The estimated curve  $\hat{\theta}_1(1 - \exp(-\hat{\theta}_2x))$  is shown as well.

Figure 4.5 shows the empirical likelihood ratio confidence regions for  $\theta$ . These contours do not look very elliptical. Some of them are not even convex.

#### 4.7 Generalized linear models

In a generalized linear model (GLM) we begin with a parametric model for the data  $Y \sim f(y; \theta)$ , where  $f$  may be either a probability density function, or a probability mass function, with a single real-valued parameter  $\theta$ . This parameter is then written as  $\theta = \tau(X'\beta)$  for predictors  $X$ , a coefficient vector  $\beta$ , and a known function  $\tau$ . The same response surface models  $X'\beta$  that are used in linear regression models may be used in GLM's. The function  $\tau$  serves, at the least, to squash the real line into the natural domain for  $\theta$ . GLM's are usually described in terms of the link function  $\tau^{-1}$  for which  $X'\beta = \tau^{-1}(\theta)$ .

Apart from normal theory regression, the most widely used GLM is logistic regression. Here  $Y_i \in \{0, 1\}$  are independent, with  $\Pr(Y_i = 1 | X_i = x_i) =$

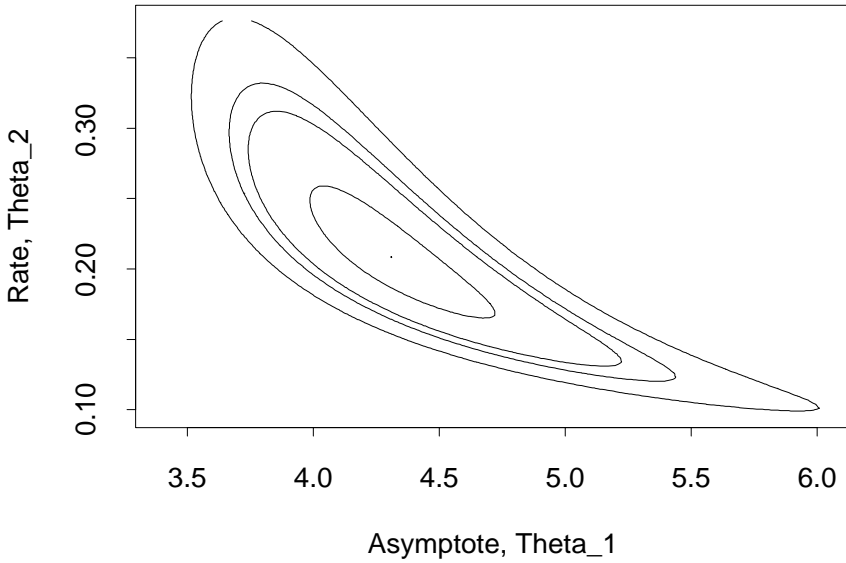


Figure 4.5 Shown are empirical likelihood contours for the parameters  $\theta_1$ , and  $\theta_2$  in the calcium uptake model. The confidence levels are 50%, 90%, 95%, and 99%.

$\tau(x'_i\beta)$ , for

$$\tau(z) = \frac{\exp(z)}{1 + \exp(z)} = \left(1 + \exp(-z)\right)^{-1}.$$

The parametric likelihood for  $\beta$  is

$$\prod_{i=1}^n [\tau(x'_i\beta)]^{Y_i} [1 - \tau(x'_i\beta)]^{1-Y_i},$$

with log likelihood

$$\sum_{i=1}^n Y_i \log \tau(x'_i\beta) + (1 - Y_i) \log(1 - \tau(x'_i\beta)),$$

and estimating equations

$$\sum_{i=1}^n \left( \frac{Y_i}{\tau(x'_i\beta)} - \frac{1 - Y_i}{1 - \tau(x'_i\beta)} \right) \tau'(x'_i\beta)x_i = 0,$$

where  $\tau'(z)$  denotes

$$\frac{d}{dz}\tau(z) = \frac{\exp(-z)}{[1 + \exp(-z)]^2} = \tau(z)(1 - \tau(z)).$$

After some algebra, the estimating equations simplify to

$$\sum_{i=1}^n x_i(Y_i - \tau(x'_i\beta)) = 0. \quad (4.14)$$

Empirical likelihood inferences for logistic regression are based on

$$\mathcal{R}(\beta) = \max \left\{ \prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i Z_i(\beta) = 0, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\},$$

where  $Z_i(\beta) = x_i(Y_i - \tau(x'_i\beta))$ .

For a generalized linear model with  $Y_i \sim f(y_i; \theta_i)$  and  $\theta = \tau(x'_i\beta)$ , the estimating equations are

$$\sum_{i=1}^n \frac{g(y_i, \tau(x'_i\beta))}{f(y_i; \tau(x'_i\beta))} \tau'(x'_i\beta) x_i = 0, \quad (4.15)$$

where as usual  $g(y, \theta) = \partial f(y; \theta) / \partial \theta$ . In practice it can pay to simplify (4.15) for the actual functions  $f$ ,  $g$ , and  $\tau$  of the model. For example, (4.14) provides insight into logistic regression that is not directly evident in (4.15).

A common source of difficulty with generalized linear models is overdispersion. The generalized linear model usually implies that the conditional variance of the response, given some predictors, is a known function of the conditional mean, given those same predictors. Overdispersed data has a conditional variance larger than what the model predicts for it. Overdispersion can invalidate statistical inferences based on parametric likelihood ratios. An empirical likelihood analysis treats the generalized linear model as a “working likelihood”, using the same maximum likelihood estimate, but substituting a more generally applicable likelihood ratio for the parametric one.

Giant cell (temporal) arteritis (GCA) is a form of vasculitis – inflammation of blood or lymph vessels. A set of data on vasculitis cases was collected in order to investigate statistical methods of separating GCA from other forms of vasculitis. There were 585 cases, with the 8 binary features recorded in [Table 4.1](#). The results of a logistic regression are shown in [Table 4.2](#) and in [Figure 4.6](#).

The coefficients  $\beta_j$  are all strongly significant except the one for scalp tenderness. For a patient with symptoms  $z_j \in \{0, 1\}$ , interest centers on the function  $\theta = \beta_0 + \sum_{j=1}^8 z_j \beta_j$  of the  $\beta_j$ . The primary interest may be in  $(1 + \exp(-\theta))^{-1}$ , the probability under the logistic regression model that this patient has GCA. [Figure 4.7](#) shows the empirical likelihood ratio for this probability for nine hypothetical patients, the  $k$ 'th one of which has the first  $k - 1$  symptoms and no others. For a patient with the first three or fewer of the symptoms, the likelihood concentrates around low probabilities of GCA. If six or more symptoms are present, then we may be similarly confident that the probability of GCA is very high. A GCA probability near 1/2 would describe great uncertainty. For a patient with the first four or five symptoms and no others, not only is the outcome uncertain, even the amount of uncertainty is not well determined from the data.

Variable	Equals 1 if and only if
Headache	New onset of localized headache
Temporal artery	Tenderness or decreased pulsation
Polymyal rheumatism	Aching and morning stiffness
Artery biopsy	Histological changes on biopsy, showing destructive inflammatory process
ESR	Erythrocyte sedimentation rate $\geq 50$ mm/hour
Claudication	Fatigue and discomfort while eating
Age	Disease onset after age 50
Scalp tenderness	Tender areas or nodules over scalp, away from arteries

Table 4.1 *Binary predictors of GCA. See the source (Bloch et al. 1990) for full definitions.*

Another quantity of interest here is the predictive accuracy of the logistic regression. Suppose that a threshold  $c$  is used so that if  $Z\beta > c$  then  $Y$  is predicted to be 1, otherwise  $Y$  is predicted to be 0. The value of  $c$  might be 0, or it might be adjusted to take account of the prior odds that  $Y = 1$ , or the ratio  $L(1, 0)/L(0, 1)$  where  $L(j, k)$  is the loss from predicting  $Y = j$  when  $Y = k$ .

Variable	Coefficient	Log likelihood
Intercept	-8.83	-272.33
Headache	1.54	-7.55
Temporal artery	2.45	-11.21
Polymyal rheumatism	1.07	-3.59
Artery biopsy	3.56	-43.96
ESR	1.69	-6.26
Claudication	2.06	-4.26
Age	3.50	-19.72
Scalp tenderness	-0.21	-0.59

Table 4.2 *Estimated logistic regression coefficients  $\hat{\beta}_j$  for GCA predictors, with empirical log likelihood for  $H_0 : \beta_j = 0$ . Scalp tenderness is not a statistically significant predictor. The others are all strongly significant.*

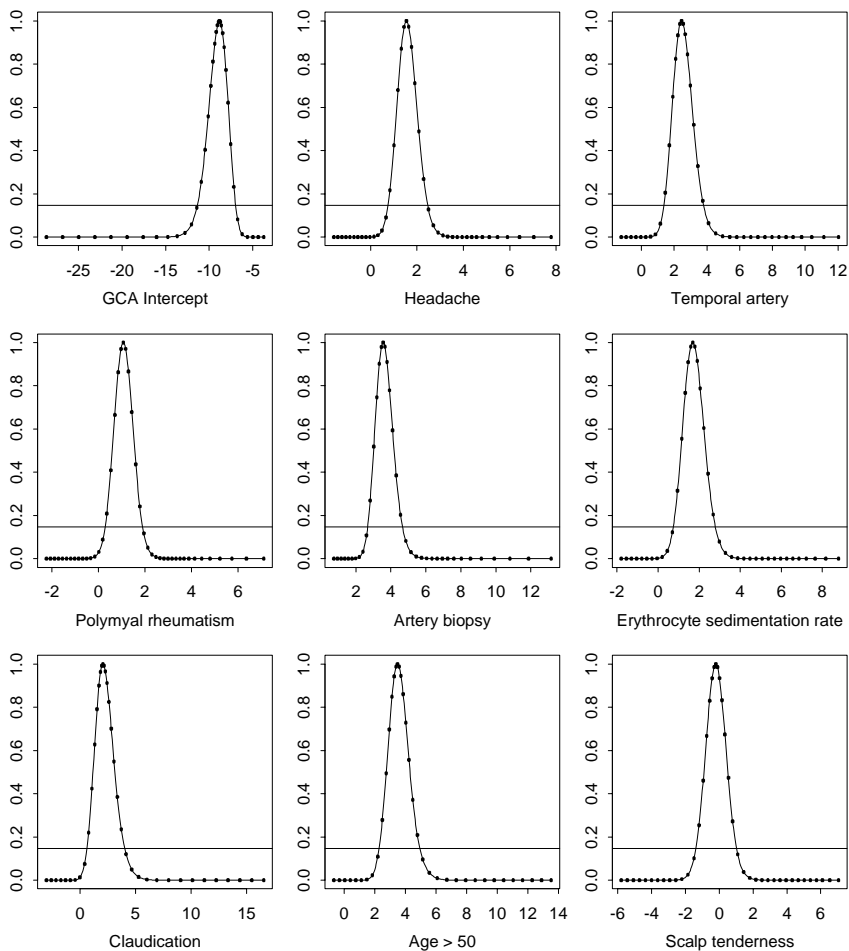


Figure 4.6 Shown are empirical likelihood functions for the 9 parameters in the logistic regression of the GCA data. The plotting limits for  $\beta_j$  correspond to  $\log \mathcal{R}(\beta_j)$  approximately equal to  $-25.0$ . The horizontal lines denote approximate 95% confidence levels, using a  $\chi_{(1)}^2$  calibration of  $-2 \log \mathcal{R}(\beta_j)$ .

Define  $\theta_0$  and  $\theta_1$  through estimating equations

$$0 = E(Y \times (1_{Z\beta \leq c} - \theta_1)), \quad \text{and}$$

$$0 = E((1 - Y) \times (1_{Z\beta \geq c} - \theta_0))$$

Then  $\theta_j$  is the probability of making a mistaken prediction, when  $Y = j$ . The

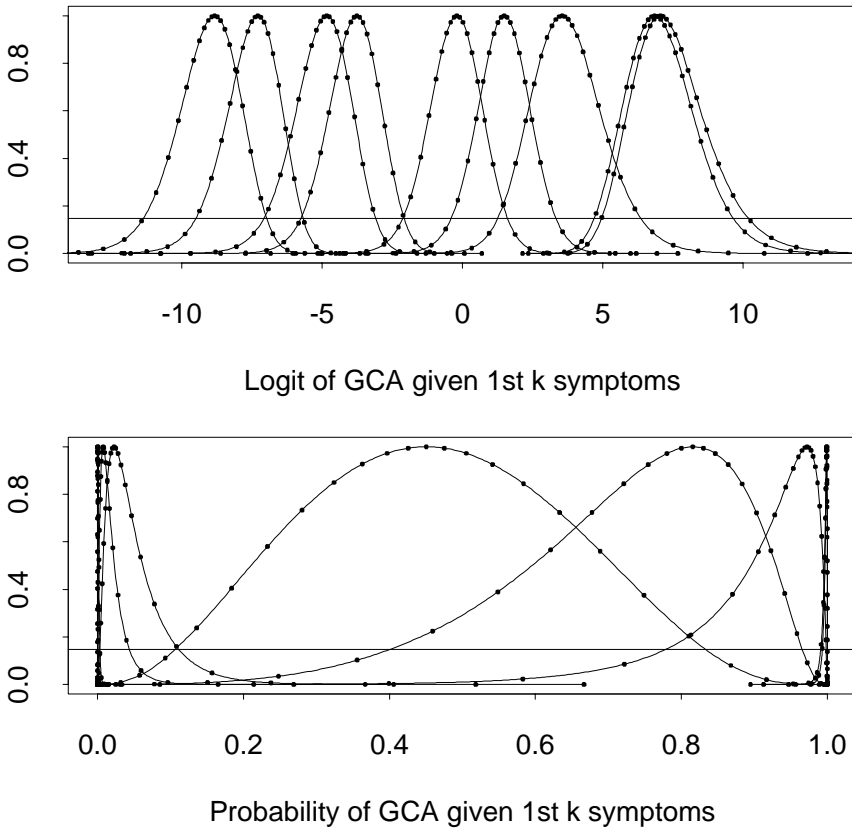


Figure 4.7 Shown are empirical likelihood functions for nine hypothetical patients. Patient  $k$  presents the first  $k - 1$  symptoms and only those, for  $k = 1, \dots, 9$ . The top figure gives the empirical likelihood ratio function for  $\sum_{j=0}^{k-1} \beta_j$ . As the number of symptoms increases from zero to nine each curve is to the right of the previous one, except that the curve for all eight symptoms is just to the left of the one for the first seven symptoms. The lower figure plots the same likelihoods versus  $(1 + \exp(-\sum_{j=0}^k \beta_j))^{-1}$ , giving a likelihood for the estimated probability of GCA.

quantity  $1 - \theta_1$  is the sensitivity of a logistic regression classifier for GCA. We will work with  $c = 0$  for simplicity.

Sample versions of  $\theta_0$  and  $\theta_1$  are subject to a bias, because the sample version of  $\beta$  has been fit to the data. In an example like this with a large number of observations and relatively few parameters, such bias is likely to be small. Thus we might consider estimating  $\theta_j$  and forming confidence intervals for it by empiri-

cal likelihood. A difficulty arises because the Heaviside function  $H(u) = 1_{u \geq 0}$  is discontinuous, making both optimization and theory much harder. We replace  $H$  by the smooth function  $G(u, \epsilon) = \Pr(t_{(4)} \leq u/\epsilon)$  where  $t_{(4)}$  is a Student's  $t$  random variable on four degrees of freedom and  $\epsilon > 0$ . Taking  $\epsilon = 0.05$ , the estimating functions are:

$$\begin{aligned} 0 &= E\left(Y \times (G(c - Z\beta, 0.05) - \theta_1)\right) \\ 0 &= E\left((1 - Y) \times (G(Z\beta - c, 0.05) - \theta_0)\right). \end{aligned}$$

The function  $G(\cdot, 0.05)$ , shown in Figure 4.8, is continuous and differentiable and is within 0.01 of  $H$  for values of  $u$  with  $\exp(u)/(1 + \exp(u))$  outside the interval  $[0.45, 0.55]$ . This substitution makes  $\theta_j$  more tractable at the cost of blurring the error count for near misses.

Figure 4.9 plots the empirical likelihood for the smoothed conditional error probabilities, when using a threshold of  $c = 0$ . The flatness at the top of these curves is not very common in profile log likelihoods. On inspection of the data, there are a number of observations with  $Y = 0$  and  $Z\hat{\beta}$  just barely less than 0. Small movements in the logistic regression parameters can produce modestly large positive  $Z\hat{\beta}$  values for these observations.

#### 4.8 Poisson regression

Poisson regression is a generalized linear model in which  $Y_i \sim \text{Poi}(\tau(x'_i\beta))$ . Here  $x_i$  is a vector of predictors, usually including a component always equal to 1. The most widely used model has  $\tau(z) = \exp(z)$ .

The number of home runs  $Y$  hit by a baseball player in one year may have approximately a Poisson distribution. It is reasonable to expect the number of home runs to depend on the number of times  $m$  the player came to bat, as well as the number of years  $t$  that the player has been playing. A natural model is that in year  $i$ ,  $Y_i \sim \text{Poi}(\lambda_i)$  where

$$\begin{aligned} \lambda_i &= \exp(\beta_0 + \beta_1 t_i + \beta_2 \log(m_i)) \\ &= m_i^{\beta_2} \exp(\beta_0 + \beta_1 t_i). \end{aligned}$$

It is natural to set  $\beta_2 \equiv 1$  in order to study the number of home runs per at bat. Such a constraint is called an offset in generalized linear modeling. Thus we consider the model

$$Y_i \sim \text{Poi}(m_i \exp(\beta_0 + \beta_1 t_i)).$$

Baseball fans tend to study at bats per home run  $m_i/Y_i$ . Figure 4.10 shows this quantity plotted against the year for two singular home run hitters of the 20th century: Babe Ruth and Hank Aaron. It is well beyond the scope of this text to attempt to compare two players from different eras. However, the coefficient  $\beta_1$  can be interpreted as a comparison for an individual player. A positive value

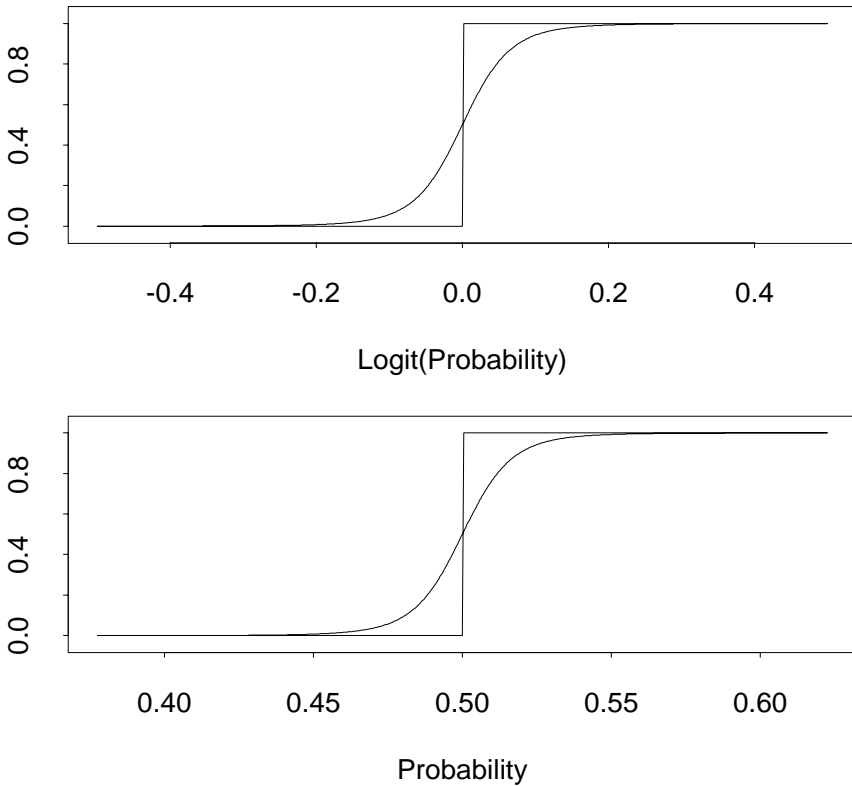


Figure 4.8 The top plot shows  $G(u, 0.05)$  and  $H(u)$  for the logit,  $u = \log(\pi/(1-\pi))$ . The bottom plot shows  $G(u, 0.05)$  and  $H(u)$  versus the probability  $\pi = \exp(u)/(1 + \exp(u))$ .

describes a player whose home run production is increasing over time. A negative value has the opposite interpretation.

It appears from the plot that Ruth's home run production was fluctuating around a decreasing trend line. Aaron's was mostly increasing, except for his final two (or possibly three) seasons. A trend in home run production could be due to changes in a player's skill, or to many other factors, such as an opposite trend in the pitching talent, or changes in baseball manufacture, stadium size, and team strategy.

For Babe Ruth,  $\hat{\beta}_1 = -.01841$ , corresponding to a drop of about 1.84% per year in home run production per at bat. For Hank Aaron,  $\hat{\beta}_1 = 0.01420$ , corresponding to an increase of about 1.42% per year.

We can investigate these trends by constructing confidence intervals for  $\beta_1$ . A least squares regression confidence interval would ignore the fact that for Poisson data the variance is equal to the mean. A confidence interval based on the Poisson

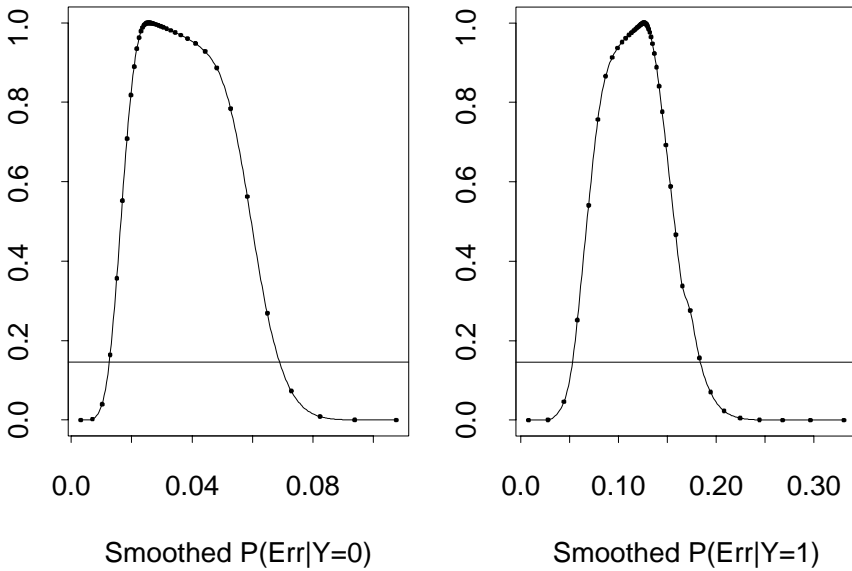


Figure 4.9 *The left curve shows the empirical likelihood function for the smoothed probability of error for a patient with GCA. The right curve shows the empirical likelihood function for the smoothed probability of error for a patient without GCA.*

likelihood could be in error because of overdispersion. Finally a confidence interval based on likelihood curvature at the MLE might be inaccurate because of the strong curvature in the exponential activation function.

Figure 4.11 shows profile empirical likelihood ratio curves for  $\beta_1$ , for these two players. The bootstrap threshold for  $-2 \log \mathcal{R}(\beta_1)$  is approximately 6.65 for Babe Ruth and 6.51 for Hank Aaron. These are indicated by horizontal calibration lines at  $\exp(-6.65/2) = 0.0360$  and  $\exp(-6.51/2) = 0.0385$ , respectively. A short vertical line through  $\beta_1 = 0$  shows that 0 is in the confidence interval for Aaron, and just barely inside the one for Ruth.

It is plausible that Ruth's home run rate only appears to decrease because of sampling fluctuations. Aaron's home run rate could more easily have been constant. Running the analysis on all of Aaron's seasons but the last two, one finds that 0 is clearly outside of the confidence interval. The interpretation of this is that Aaron might have been steadily increasing in home run output for the first 21 of his 23 seasons, but that the last two seasons do not fit the linear model.

## At bats per home run

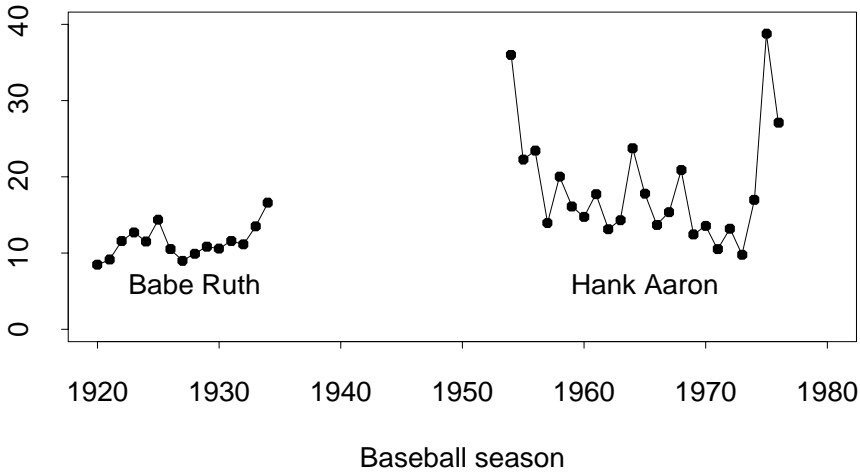


Figure 4.10 Shown are the number of at bats per home run for Babe Ruth and Hank Aaron, in each year of their careers. The left curve is for Babe Ruth, the right is for Hank Aaron. The Aaron data can be found at <http://Baseball-Reference.com> and the Ruth data is from Stapleton (1995).

### 4.9 Calibration, prediction, and tolerance regions

In linear regression, a common problem is to construct a confidence interval for  $\beta_0 + \beta_1 x_0$ , where  $x_0$  is a specified value of the predictor variable. A related problem, called calibration or inverse regression, is to find a confidence region for the value  $x_0$  with  $\beta_0 + \beta_1 x_0 = y_0$ , for a given response value  $y_0$ .

These problems are similar in that they may be handled by maximizing likelihood, parametric or empirical, subject to the additional constraint

$$\beta_0 + \beta_1 x_0 = y_0.$$

For regression,  $x_0$  is fixed and the likelihood ratio is found for each  $y_0$ , after maximizing over  $\beta_0$  and  $\beta_1$ . For calibration,  $x_0$  varies for a given  $y_0$ . There is an important practical difference: when the slope  $\beta_1$  is not well determined the confidence set for  $x_0$  is not necessarily a finite interval. It can be the whole real line or even the set theoretic complement of a finite interval.

Prediction intervals extend easily to nonlinear and generalized linear models. Calibration intervals are more complicated. Specifying  $y_0$  imposes only one constraint on  $x_0$ , so if the predictor is in a  $p$ -dimensional space, there will generally be a  $p - 1$  dimensional space of  $x_0$  values consistent with  $y_0$  at the MLE of the

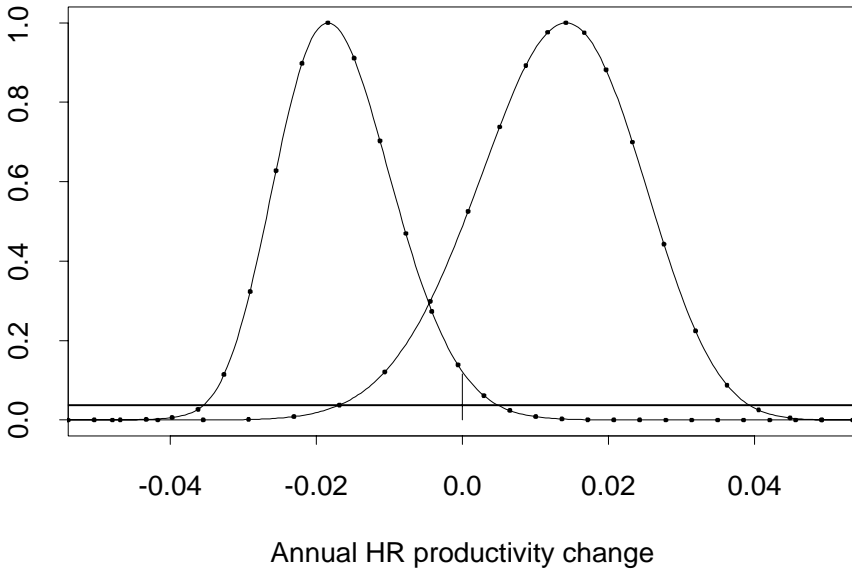


Figure 4.11 Shown are empirical likelihood ratio curves for the time coefficient  $\beta_1$  in the home run intensities of Babe Ruth and Hank Aaron. The coefficient  $\beta_1$  measures a player's annual improvement in home run productivity, as described in the text. The curve for Ruth is to the left of that for Aaron. Horizontal reference lines and a short vertical reference line show that 0 is in the confidence interval for  $\beta_1$  for both players, as described in the text.

parameter. More generally, when the dimension of  $X$  is larger than that of  $Y$ , we can ordinarily expect that calibration estimating equations are underdetermined, while if  $Y$  has the larger dimension, we might anticipate the calibration problem to be overdetermined.

A related problem is that of tolerance intervals or regions. Given  $x_0$ , we seek a set that with confidence  $1 - \alpha$  contains at least  $1 - \gamma$  of the probability in the distribution of corresponding  $Y_0$  values. Consider a linear regression problem with observations  $(X_i, Y_i) \in \mathbb{R}^2$  independent and identically distributed for  $i = 1, \dots, n$ . The estimating equations

$$0 = E(Y - \beta_0 - \beta_1 X) \quad (4.16)$$

$$0 = E(X(Y - \beta_0 - \beta_1 X)) \quad (4.17)$$

$$0 = E((Y - \beta_0 - \beta_1 X)^2 - \sigma^2) \quad (4.18)$$

$$0 = E(1_{Y < \beta_0 + \beta_1 X + \tau \sigma} - 0.95) \quad (4.19)$$

define the regression intercept, slope, and error standard deviation to be  $\beta_0$ ,  $\beta_1$ , and  $\sigma$ , respectively. They also define  $\tau$  such that 5% of the  $Y$  values lie  $\tau$  or more

standard deviations above the regression line. For a tolerance interval,  $\beta_0 + \beta_1 x_0 + \tau\sigma$  is of interest, while  $\beta_0, \beta_1, \tau,$  and  $\sigma$  themselves are nuisance parameters.

Equation (4.19) above is not differentiable with respect to the parameters, and so Theorem 3.6 does not apply. Theorem 3.4 does apply if we are satisfied with a joint confidence region  $C^{1-\alpha}$  for  $(\beta_0, \beta_1, \sigma, \tau)$ . We could construct a region  $C^{1-\alpha}$  using a  $\chi^2_{(4)}$  calibration. The set  $\{\beta_0 + \beta_1 x_0 + \tau\sigma \mid (\beta_0, \beta_1, \sigma, \tau) \in C^{1-\alpha}\}$  then has asymptotic coverage greater than or equal to the nominal level from the  $\chi^2_{(4)}$  calibration. Intuitively we expect that the right coverage level should be obtained using one degree of freedom. Presently known results do not let us conclude this in general. See Chapter 10.6 for some related results.

#### 4.10 Euclidean likelihood for regression and ANOVA

Applying the Euclidean log likelihood to the regression problem produces a test statistic

$$(\hat{\beta} - \beta_0)' \left[ (\mathcal{X}'\mathcal{X})^{-1} \left( \sum_{i=1}^n (Y_i - X_i\hat{\beta})^2 X_i X_i' \right) (\mathcal{X}'\mathcal{X})^{-1} \right]^{-1} (\hat{\beta} - \beta_0)$$

where  $\mathcal{X} = (X_1 \ X_2 \ \dots \ X_n)'$ .

This is equivalent to using White's heteroscedasticity robust estimate

$$(\mathcal{X}'\mathcal{X})^{-1} \left( \sum_{i=1}^n (Y_i - X_i'\hat{\beta})^2 X_i X_i' \right) (\mathcal{X}'\mathcal{X})^{-1} \quad (4.20)$$

of the covariance matrix of  $\hat{\beta}$ . The true covariance of  $\hat{\beta}$  is

$$(\mathcal{X}'\mathcal{X})^{-1} \left( \sum_{i=1}^n \sigma_i^2 X_i X_i' \right) (\mathcal{X}'\mathcal{X})^{-1},$$

and so (4.20) can be thought of as using  $(Y_i - X_i'\hat{\beta})^2$  as an estimate of  $\sigma_i^2$ . For each individual  $\sigma_i^2$ , this is a poor estimate, but the  $n$  estimation errors tend to cancel and the result gives properly calibrated inferences on  $\beta$  that do not require equal variance for the observations.

If there is a common value  $\sigma_i^2 = \sigma^2$ , the true covariance simplifies to the familiar form  $\sigma^2(\mathcal{X}'\mathcal{X})^{-1}$ , for which the usual estimate is of the form  $\hat{\sigma}^2(\mathcal{X}'\mathcal{X})^{-1}$ .

For the Euclidean log likelihood the two approaches to ANOVA in Chapter 4.4 do not in general provide the same answer. Taking the  $k$  distributions approach, with distance function  $\sum_{i=1}^k \sum_{j=1}^{n_i} (n_i v_{ij} - 1)^2$  the resulting test statistic is

$$\sum_{i=1}^k n_i \frac{(\bar{Y}_{i\cdot} - \tilde{Y}_{\bullet\bullet})^2}{s_i^2},$$

where

$$s_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2,$$

and

$$\tilde{Y}_{\bullet\bullet} = \frac{\sum_{i=1}^k n_i \bar{Y}_{i\bullet} / s_i^2}{\sum_{i=1}^k n_i / s_i^2}.$$

The test statistic for regression, and the test statistic for ANOVA (with minor modifications) have both been used in applications before, as described in the bibliographic notes in Chapter 4.11.

#### 4.11 Bibliographic notes

Empirical likelihood for regression was considered by Owen (1991), who used the normal estimating equations to study first order coverage properties, including the case of nonrandom predictors. Song Chen investigated higher order terms, establishing Bartlett correctability and exhibiting the Bartlett factor, again for nonrandom predictors. Chen (1993) considers confidence regions for the regression coefficients and Chen (1994b) looks at confidence intervals for linear combinations of regression coefficients.

The formulation of empirical and Euclidean likelihoods for ANOVA is taken from Owen (1991). Jing (1995b) and Adimari (1995) independently considered the problem of comparing the means of two populations. Both find a chisquared limit for the difference in means, using a product of one sample empirical likelihoods. Adimari (1995) obtains a noncentral chisquared distribution under alternatives  $O(n^{-1/2})$  from the true difference in means. Jing (1995b) shows Bartlett correctability.

The conservatism of empirical likelihood for fixed predictors and a misspecified regression model is explained in Owen (1991). Davidian & Carroll (1987) consider variance modeling in regression problems.

La Rocca (1998) studies the coverage error of empirical likelihood for linear regression models, including several error distributions, and both equal and unequal error variances. Bootstrap calibration of empirical likelihood proves to be most reliable.

The most basic form of model selection for regressions is conducted by fitting the model with and without one of the predictor variables and comparing the sum of squares. In empirical likelihood, the analogous procedure is to fit the model with and without constraining the corresponding parameter to be zero, and to compare the constrained and unconstrained empirical log likelihood ratios. Constraining a parameter to zero is different from dropping the corresponding predictor, because the constrained fit keeps the residuals uncorrelated with the missing predictor. Based on the results in Qin & Lawless (1994) we would expect this extra information to be helpful. Kolaczyk (1995) investigated empirical likelihood

model selection, introducing an empirical information criterion (EIC) analogous to AIC (Akaike's information criterion) given by Akaike (1973).

Generalized linear models were introduced by Nelder & Wedderburn (1972). McCullagh & Nelder (1983) is the standard reference on GLM's. Empirical likelihood for GLM's was investigated by Kolaczyk (1994), who considered GLM's with a fixed amount of overdispersion. He also considered estimating an overdispersion constant and modeling the overdispersion via a link and a linear model. The usual inferences for GLM's do not require that the motivating parametric model hold. They can instead be defined through a quasi-likelihood. See Wedderburn (1974). This weakens the model to a requirement that the conditional variance of  $Y$  given  $X = x$  be functionally related to the conditional mean, that is  $\sigma^2(x) = h(\mu(x))$  for some known function  $x$ . There is no reason in general to expect data that depart from the parametric model to satisfy the quasi-likelihood condition, although the introduction of an overdispersion parameter or model can mitigate the problem. Nelder & Pregibon (1987) introduce extended quasi-likelihood to model overdispersion. See also Jorgensen (1987). Efron (1986) introduces double exponential families.

The GCA data is from Bloch et al. (1990). They investigate numerous prediction rules for GCA and find that logistic regression performs best. They do not use a zero threshold. Their logistic regression coefficients are not the same as the ones in Table 4.2, though they are qualitatively similar. There may have been slight differences in the data, or they might not have used maximum likelihood estimates. The coefficients in Table 4.2 agree with the ones computed by the `glm` function in S-PLUS.

The idea to study baseball home runs comes from Stapleton (1995), who provides the Babe Ruth data. Data for other players is easily available over the Internet at numerous sites. The site [Baseball-Reference.com](http://Baseball-Reference.com) contains a lot of baseball data.

Background material on nonlinear least squares asymptotics and algorithms can be found in Bates & Watts (1988) and Seber & Wild (1989). The calcium uptake data is from Rawlings (1988). Davison & Hinkley (1997) present a bootstrap analysis of it.

The estimate (4.20) is due to White (1980).

For the analysis of variance (with  $d = 1$ ), the Euclidean likelihood is essentially equivalent to the statistic used by James (1951) to test differences among group means when the variances are thought to be unequal. James (1951) used  $n_i - 1$  instead of  $n_i$  in defining  $s_i^2$ . His critical value was not taken from a  $\chi^2$  distribution but instead took account of the differing values of  $s_i^2$ .

## 4.12 Exercises

**Exercise 4.1** Consider regression with a predictor  $X \in \mathbb{R}^d$  and a response  $Y \in \mathbb{R}$ . The data are IID  $(X, Y)$  pairs, including numerous ties among the  $X$ 's. The data can then be labeled as  $X_i \in \mathbb{R}^d$ ,  $i = 1, \dots, k$ , with which we observe  $Y_{ij} \in$

$\mathbb{R}$ ,  $j = 1, \dots, n_i$ . Define  $Z_{ij}(\beta) = X_i'(Y_{ij} - X_i'\beta)$ . Let  $\bar{Y}_i = (1/n_i) \sum_{j=1}^{n_i} Y_{ij}$ ,  $\bar{Z}_i(\beta) = (1/n_i) \sum_{j=1}^{n_i} Z_{ij}(\beta)$ , and

$$s_i^2 = \begin{cases} \frac{1}{n_i-1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2, & n_i > 1 \\ 0, & n_i = 1. \end{cases}$$

Define pure error and lack of fit variances  $\sigma_{PE}^2 = \iint (y - \mu(x))^2 dF_{X,Y}(x, y)$  and  $\sigma_{LF}^2 = \int (\mu(x) - x'\beta)^2 dF_X(x)$ , where  $\mu(x) = \int y dF_{Y|X=x}(y)$ . For  $n_i > 1$ ,  $E(s_i^2) = \sigma_{PE}^2$ . Consider the estimating equations

$$\begin{aligned} 0 &= \sum_{i=1}^k w_i n_i \bar{Z}_i(\beta) \\ 0 &= \sum_{i=1}^k w_i [n_i \bar{Z}_i^2 - n_i \sigma_{LF}^2 - \sigma_{PE}^2] \\ 0 &= \sum_{i=1}^k w_i (n_i - 1) [s_i^2 - \sigma_{PE}^2]. \end{aligned}$$

Show that the quantities being averaged have expectation 0 if  $\beta$  is the least squares coefficient vector and  $\sigma_{PE}^2$  and  $\sigma_{LF}^2$  are the correct values. Do empirical likelihood confidence regions for these parameters have the correct calibration as  $k \rightarrow \infty$ ? What conditions if any are required on  $n_i$ ? Now suppose that  $k$  is fixed and that  $n_i \rightarrow \infty$ . Are empirical likelihood regions correctly calibrated? If not, suggest an alternative formulation of the problem.

**Exercise 4.2** For the cancer data, construct the empirical likelihood ratio function for the slope  $\beta_1$ , using the estimating equation  $E((C_i - \beta_1 P_i)P_i) = 0$ . Compare it to the likelihood ratio curve for  $\beta_1$  from ordinary linear regression and from regression through the origin with  $N(0, \sigma^2)$  errors.

**Exercise 4.3** Find a joint distribution for  $X \in \mathbb{R}^d$  and  $Y \in \mathbb{R}$ , and a vector  $\beta \in \mathbb{R}^d$  such that  $E(\|X\|^2(Y - X'\beta)^2) < \infty$ , but at least one of  $E(\|X\|^4) = \infty$  or  $E(\|X\|^2 Y^2) = \infty$  holds.

**Exercise 4.4** Consider a regression of  $Y_i$  on predictors  $X_i = (1, U_i, V_i)'$ ,  $i = 1, \dots, n$ . Suppose that  $n = 100$  and that  $U_{62}, U_{84}, V_{17}, V_{38}$ , and  $V_{62}$  are missing, while all other observations are available. In addition to the regression parameter vector  $\beta$  introduce a parameter for each of the missing values. The estimating equations are still

$$\sum_{i=1}^n w_i X_i(Y_i - X_i'\beta) = 0,$$

except that there are now eight parameters  $(\beta_1, \beta_2, \beta_3, U_{62}, U_{84}, V_{17}, V_{38}, V_{62})$  for these three estimating equations. Which, if any, of these parameters has a

unique NPMLE? Assume that the matrix consisting of all completely observed  $X_i$  vectors has full rank.

**Exercise 4.5** Consider the ratio

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y}_{..})^2}{\frac{1}{N-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}.$$

Suppose that  $k = 2$ ,  $n_1 = n$ , and  $n_2 = n^2$ . Consider the limit as  $n \rightarrow \infty$ . Assume that independent  $Y_{ij} \sim F_i$  have mean  $\mu_i$ , variance  $\sigma_i^2$ , and  $E(Y_{ij}^4) < \infty$ . Show that  $F$  has a  $\chi^2_{(1)}$  limit if  $\mu_1 = \mu_2$  and  $\sigma_1^2 = \sigma_2^2 > 0$ . What happens if the means are equal but the variances are not?

**Exercise 4.6** Verify that the heteroscedastic regression estimating equations (4.8) and (4.9) are the likelihood equations for the model  $Y \sim N(X'\beta, \exp(2Z'\gamma))$ .

**Exercise 4.7** Consider these two ways of computing

$$\begin{aligned} \tau(z) &= \frac{\exp(z)}{1 + \exp(z)} \\ &= [1 + \exp(-z)]^{-1}, \end{aligned}$$

the squashing function for logistic regression. The IEEE floating point systems have numbers that represent  $\pm\infty$ . As one would expect,  $1/\infty = 0$  and  $1/0 = \infty$ . There is even a reciprocal  $-0$  for  $-\infty$  with  $-0$  equaling  $0$ . These systems also have values NaN that designate “not a number”. A floating point value of NaN arises from operations like dividing  $0$  by  $0$ , subtracting  $\infty$  from  $\infty$ , multiplying  $0$  by  $\infty$  or dividing  $\infty$  by  $\infty$ .

The first expression for  $\tau$  above can produce NaN for finite  $z$ , the second one should not produce NaN for any  $z \in [-\infty, \infty]$ . Find two ways of expressing  $\tau'(z) = d\tau(z)/dz$  for which one way produces NaNs and the other does not. Assume that  $\exp(\infty) = \infty$  and  $\exp(-\infty) = 0$ . Find a way of computing  $\tau''(z) = d^2\tau(z)/dz^2$  for which NaNs are not produced for any  $z \in [-\infty, \infty]$ .

**Exercise 4.8** The specificity of a classifier for predicting that  $Y = 1$  is the probability that  $Y = 1$  given that the classifier predicts  $Y = 1$ . There is commonly a trade-off between sensitivity and specificity. For a logistic regression predicting  $Y = 1$  when  $X\beta > c$ , the tradeoff is governed by the choice of  $c$ . Write the estimating equation for the specificity of a logistic regression rule. Is it smooth in the parameters?