# Overdispersion Diagnostics for Generalized Linear Models

DIANE LAMBERT
*AT&T Bell Laboratories*
*Murray Hill, NJ 07974*
and
KATHRYN ROEDER *
*Department of Statistics*
*Yale University*
*New Haven, CT 06520*

**Abstract**

Generalized linear models (GLMs) are simple, convenient models for count data, but they assume that the variance is a specified function of the mean. Although overdispersed GLMs allow more flexible mean – variance relationships, they are often not as simple to interpret nor as easy to fit as standard GLMs. This paper introduces a *convexity plot*, or *C-plot* for short, that detects overdispersion and *relative variance curves* and *relative variance tests* that help to understand the nature of the overdispersion. Convexity plots sometimes detect overdispersion better than score tests, and relative variance curves and tests sometimes distinguish the source of the overdispersion better than score tests.

**Keywords** Mixture, random coefficient, residuals, score tests, variance inflation

## 1 INTRODUCTION

Convenient generalized linear models (GLMs) for count data, such as logistic regression and log-linear Poisson regression, require the variance to be a known function of the mean. But count data are often more heterogeneous than the GLM variance allows. Quasi-likelihood methods relax the mean–variance relationship slightly by allowing the variance to be inflated by a constant (Wedderburn 1974). Extended quasi-likelihood methods and double exponential families (*DEF*) relax the relationship further by allowing the variance multiplier to depend on covariates (Nelder and Pregibon 1987; Efron 1986). Overdispersion can also be

*Kathryn Roeder was supported by NSF grants DMS9201211 and DMS9257006.

accommodated by randomizing the mean of the GLM to obtain beta-binomial and gamma-Poisson models, among others (*e.g.*, Cox 1983). Often, there is no physical justification for extended quasi-likelihood, DEF, or random mean GLM models, however.

In contrast, binomial and Poisson models with random coefficients try to represent the source of overdispersion explicitly. For example, in a random coefficient log-linear Poisson regression, the response $Y$ is Poisson($e^{\alpha+x\beta}$) given $\alpha$ and $\beta$, but each individual has a different random baseline $\alpha$ or different responsiveness to treatment $\beta$ (*e.g.*, Pierce and Sands 1975; Breslow 1984; Follmann and Lambert 1989, Butler and Louis 1992). Random coefficients may represent inter-subject variability, errors-in-variables (Stefanski, Carroll,and Ruppert 1986), or unmeasured covariates. For instance, if the true log-mean is $\alpha + x\beta + z\gamma$ but only $x$ is recorded, then the assumed log-mean $\alpha^* + x\beta$ has a random intercept $\alpha^* = \alpha + z\gamma$ that varies with $z$. If the true log-mean is $\alpha + x\beta + xu\gamma$ where $u$ is a variable that is uncorrelated with $x$ and the interaction $xu$ is omitted, then the assumed log-mean $\alpha + x\beta^*$ has random slope $\beta^* = \beta + u\gamma$.

Under all these models – extended quasi-likelihood, DEF, random mean GLM and random coefficient GLM – the extent of the overdispersion depends on $x$. Letting the overdispersion depend on $x$ is appealing, but the resulting models are not as simple to specify, fit, or interpret as standard GLMs. In particular, the mean of a random coefficient GLM is not the GLM mean evaluated at the average of the random coefficients (see, for example, Neuhaus *et al.* 1991 and Theorem 4 below). Thus, without fitting any of these models, we would like to know whether there is serious overdispersion, and, if there is, which overdispersed model should be fit.

This paper introduces a *convexity plot*, or *C-plot* for short, that detects overdispersion in GLMs (Section 3) and *relative variance curves* and *relative variance tests* that help to identify the nature of the overdispersion (Section 4). Although there are score tests for detecting DEF (Smyth 1989 and Ganio and Schafer 1992), random mean (Smith and Heitjin 1993), and random coefficient (Bartoo and Puri 1967, Zelterman and Chen 1988, and Dean 1992) GLMs, there is still a place for $C$-plots and relative variance curves and tests. Section 3 shows that $C$-plots detect some kinds of overdispersion better than score tests do. Section 4 shows that relative variance curves and tests sometimes discern the nature of the overdispersion better than locally most powerful score tests do. Examples of $C$-plots and relative variance curves are given in Section 2. Proofs are given in the Appendix.

It is perhaps important to note that our tools do not distinguish lack-of-fit (incorrect mean) from true overdispersion (incorrect variance). We assume throughout that the mean has already been specified as well as possible and any remaining poor fit must be attributed to overdispersion.

## 2 AN EXAMPLE

Efron (1978, 1986) presents data on the standardized annual rainfall $x$ and the fraction of subjects $y/m$ who tested positive for toxoplasmosis in 34 cities in El Salvador. As Efron notes, a cubic logistic regression fits the data better than a linear or quadratic logistic regression, but even the cubic does not fit the data well. Including an overdispersion factor that depends on $m$ seems to improve the fit (Efron 1986), but Ganio and Schafer (1992) found that score tests do not support modeling overdispersion as a function of $m$ or $x$.

Here we consider two simpler models for the toxoplasmosis data: a signed quadratic logistic regression model with $logit(p) = \alpha + \beta_1 x + \beta_2 \, sign(x) \, x^2$ and a linear logistic regression model. The signed quadratic regression fits the data almost as well as the cubic regression does (the cubic is worse for moderate $x$ and better for large $x$). The linear model does not fit as well, but the other models are non-monotone and their biological interpretation is unclear. Thus, we ask whether there is serious overdispersion from the linear model and, if so, whether the signed quadratic term removes the overdispersion.

The convexity or $C$-plot for a logistic regression shows the average of the relative likelihood of $p$ and the estimated probability $\hat{p}_i$ from the logistic regression as a function of $p$. That is,

$$C(p) = n^{-1} \sum_{i=1}^{n} \left( \frac{p}{\hat{p}_i} \right)^{y_i} \left( \frac{1-p}{1-\hat{p}_i} \right)^{m_i - y_i}$$

is plotted against $p$. (See Section 3 for details.) The more convex $C(p)$ appears, the more evidence there is of overdispersion or an omitted variable. Figure 1 shows that for the toxoplasmosis data $C(p)$ is convex for the linear logistic model, but it is not convex for the signed quadratic model. The next question is whether the overdispersion in the linear logistic model varies with $x$ or $m$ or neither.

Figure 1 about here.

The relative variance curves $\hat{R}_0(m)$ and $\hat{R}_1(x)$ are nonparametric estimates of the effects of $m$ and $x$ on an average ratio of the true variance of $y$ to the estimated variance under the GLM. (See Section 4 for details.) If there is no overdispersion, $\hat{R}_0(m)$ and $\hat{R}_1(x)$ are nearly zero. If the overdispersion can be modeled as a random intercept, then $\hat{R}_0(m)$ and $\hat{R}_1(x)$ are nearly positive constants. If $x$ has a random slope, then $\hat{R}_1(x)$ is a positive quadratic.

Figure 2 shows the relative variance curves for the toxoplasmosis data under the linear logistic regression. The points are adjusted residuals from an estimated model of the ratio of the true variance of $y$ to the GLM variance. Because $\hat{R}_1(x)$ is nearly quadratic and $\hat{R}_0(m)$ is more nearly zero, the linear logistic fit is more likely to be improved by introducing a random slope or an interaction between $x$ and an uncorrelated variable $z$ than it is to be improved by modeling the overdispersion as a function of $m$. As we have already seen, adding the product of $x$ and $|x|$ removes the overdispersion found by the $C$-plot.

Figure 2 about here.

## 3 CONVEXITY PLOT

### 3.1 The Concept

A random variable $Y$ has a one parameter exponential family density $f$ if it is the sum of $m$ independent random variables and it can be written as

$$f(y;\eta) = e^{y\eta-\psi(\eta)} \cdot d\,\gamma(y); \tag{1}$$

$\eta$ is the natural parameter and $\gamma$ is a supporting measure. The mean and variance of $Y$ are $m\mu = \psi'(\eta)$ and $V(\mu) = \psi''(\eta)$. For a Poisson, $m = 1$.

Under a GLM, $Y_1, \ldots, Y_n$ are independent with exponential family density $f$ and for a vector $x_i = (x_{i1}, \ldots, x_{ip})^t$ of $p$ known explanatory variables and an arbitrary increasing function $h$ called a *link*, $h(\mu_i) = \alpha + x_i^t\beta$ (see McCullagh and Nelder 1989). Under the *canonical link* $l$, $l(\mu_i)$ is the natural parameter of $f$. The canonical links for the binomial and Poisson are the *logit* and *log* respectively. To avoid some technical details, we assume throughout that the covariates have been scaled to lie between -1 and 1.

**Definition 1** *The* convexity plot *or* $C$-plot *for a GLM with its canonical link $l$ plots $C(\mu)$ against $\mu$ where*

$$C(\mu) = n^{-1}\sum_{i=1}^{n} \frac{f(y_i; l(\mu))}{f(y_i; \hat\alpha + x_i^t\hat\beta)} = n^{-1}\sum_{i=1}^{n} C_i(\mu)$$

*and $\hat\alpha + x_i^t\hat\beta$ is the maximum likelihood estimate* (mle) *of $\alpha + x_i\beta$. For any increasing link $h$,*

$$C(\mu) = n^{-1}\sum_{i=1}^{n} \frac{f(y_i; l(\mu))}{f(y_i; l(h^{-1}(\hat\alpha + x_i^t\hat\beta)))}.$$

For logistic regression

$$C(p) = n^{-1}\sum_{i=1}^{n} \left(\frac{p}{\hat p_i}\right)^{y_i} \left(\frac{1-p}{1-\hat p_i}\right)^{m_i-y_i},$$

where $\hat p_i = exp(\hat\alpha + x_i^t\hat\beta)/(1 + exp(\hat\alpha + x_i^t\hat\beta))$ and $p \in (0,1)$. For log-linear Poisson regression,

$$C(\mu) = n^{-1}\sum_{i=1}^{n} e^{\hat y_i - \mu} \left(\frac{\mu}{\hat\mu_i}\right)^{y_i},$$

where $\hat\mu_i = exp(\hat\alpha + x_i^t\hat\beta)$ and $\mu > 0$.

Shaked (1980) proved that each summand $C_i$ is convex in $\mu$ under mixing if the true mean $h^{-1}(\alpha + x_i^t\beta)$ is substituted for $h^{-1}(\hat\alpha + x_i^t\hat\beta)$. Lindsay (1986) extended Shaked's result to estimated means. Later, Lindsay and Roeder (1992) summed the terms with estimated means and subtracted 1 to develop a *homogeneity gradient* that detects mixing for identically distributed responses. When there are no covariates, $C(\mu) - 1$ is exactly their homogeneity gradient. We do not use the term homogeneity gradient here because $C(\mu)$ is not a derivative of most overdispersed GLMs in any obvious sense.

It is the shape of the $C$-plot that suggests the presence of overdispersion. Theorem 2 shows that $C(\mu)$ tends to be convex when the data are from a random mean GLM, random coefficient GLM, or DEF model. Theorem 3 shows that $C(\mu)$ also tends to be convex under overdispersed DEF models. In contrast, the $C$-plot is unlikely to be convex under a GLM, because the log-likelihood for a GLM with its canonical link or many other standard links is approximately concave.

**Theorem 2** *If $Y_1, \ldots, Y_n$ are from a GLM with random coefficients or random means, then $C(\mu)$ converges to a convex function on bounded intervals of $\mu$ as $n$ increases.*

**Theorem 3** *For bounded intervals of $t$, if $Y_1, \ldots, Y_n$ are from an overdispersed DEF regression model and $(\hat{\alpha}, \hat{\beta})$ is consistent for $(\alpha, \beta)$, then $C(\mu)$ converges to a convex function as $n$ increases.*

## 3.2 An Improved $C$-plot

Although Theorems 2 and 3 show that $C(\mu)$ should be approximately convex on bounded intervals of $\mu$ if $Y_1, \ldots, Y_n$ are from an overdispersed GLM and $n$ is large, in practice there is a complication. Because each summand $C_i$ approaches its minimum value of 0 when $\mu$ moves far from $\hat{\mu}_i$, there may be no range of $\mu$ for which all the summands $C_i$ are convex if the $\hat{\mu}_i$'s are far apart. To keep $C_i$ more nearly convex in its tails when it is convex near $\hat{\mu}_i$, we extend its tails linearly. The convexity plot then shows

$$C(\mu) = n^{-1} \sum_{i=1}^{n} D_i(\mu)$$

against $\mu$ where

$$D_i(\mu) = \begin{cases} C_i(t_{1i}) + (\mu - t_{1i}) C_i{}'(t_{1i}) & \text{if } \mu < t_{1i} \\ C_i(\mu) & \text{if } t_{1i} \leq \mu \leq t_{2i} \\ C_i(t_{2i}) + (\mu - t_{2i}) C_i{}'(t_{2i}) & \text{if } \mu > t_{2i}, \end{cases}$$

and $t_{1i} = \hat{\mu}_i - k\hat{\sigma}_i$, $t_{2i} = \hat{\mu}_i + k\hat{\sigma}_i$ for a positive constant $k$ and $\hat{\sigma}_i^2 = V(\hat{\mu}_i)/m_i$. For logistic regression

$$C_i{}'(\mu) = \left( \frac{y_i}{\mu} - \frac{m_i - y_i}{1 - \mu} \right) C_i(\mu)$$

and for log-linear Poisson regression

$$C_i{}'(\mu) = \left( \frac{y_i}{\mu} - 1 \right) C_i(\mu).$$

Often, the choice of $k$ is unimportant, because $C$ has the same shape for a wide range of $k$. For example, the $C$-plots in Figure 1 have the same shape for all $k$.

A final complication is that convexity can be difficult to see. Sometimes, it helps to plot the second derivative $C''(\mu)$ along with $C(\mu)$. If $C''(\mu) \geq 0$ for all $\mu$, there is evidence of overdispersion. The tails of $C_i''(\mu)$ may need to be modified, however, because $C_i''(\mu)$ approaches zero when $\mu$ moves away from $\hat{\mu}_i$. If so, we replace $C_i''(\mu)$ with $C_i''(t_{1i})$ and $C_i''(t_{2i})$ beyond $t_{1i}$ and $t_{2i}$ respectively. Because of the linearization of the tails of $C_i(\mu)$, the convexity check can be less useful in the tails of the range of the $\mu_i$'s.

The $C$-plot is intended to be a simple diagnostic, and hence we do not try to quantify the extent of convexity. If a $C$-plot looks convex and its second derivative $C''(\mu)$ is positive (except possibly in its tails), then we would explore fitting overdispersed models. As the simulations below show, convexity is often visually obvious.

If convexity of a score plot must be quantified, the score test statistics for random coefficients, which were introduced by Bartoo and Puri (1967) and described by Zelterman and Chen (1988) and Dean (1992), can be interpreted as numerical summaries of the extent of convexity. The score test for a random intercept is proportional to $\sum_i C_i''(\hat{\mu}_i)$ while the score test for a random slope at the $j$'th coefficient is proportional to $\sum_i x_{ij} C_i''(\hat{\mu}_i)$. Both tests reject for large positive values. That is, both tests are equivalent to checking for convexity of the $C$-plot at the estimated means. As will be seen in Section 3.3, the frequency with which the C-plot is convex and the rate at which the random coefficient score tests reject are about equal under local random coefficient models. $C$-plots and random coefficient score tests are not redundant, however. Under random coefficient alternatives, there are samples for which the $C$-plot detects overdispersion and the score test fails to reject, and other samples for which the score tests reject and the $C$-plot is not convex. Furthermore, the $C$-plot has been designed to detect non-local alternatives that score tests do not detect well. (For a non-local example, see the contaminated Poisson simulation in Section 3.3.1.)

## 3.3 Simulated Behavior of the $C$-plot

Several simulations with 2000 runs each were carried out in $S$ (Becker, Chambers and Wilks, 1988). (That reference includes information on the random number generators.) In each run, there were $n = 50$ responses. The values of the one covariate $x$ were generated anew in each run according to a uniform$(-1, 1)$ distribution. Section 3.3.1 considers $C$-plots of log-linear and overdispersed log-linear Poisson regressions; Section 3.3.2 considers $C$-plots of logistic and overdispersed logistic regressions.

Each simulation produced 2000 $C$-plots computed on a fixed grid of equally spaced points. These $C$-plots include a range of shapes. For example, under an overdispersed model most of the 2000 $C$-plots may be convex, many may be concave, and a few may have peculiar, intermediate shapes resembling sideways S's. Such a range of shapes is unavoidable, because tests and diagnostics rarely have power one. But it is hard to summarize a set of curves with such different shapes. As curves change shape from convex to concave, however, their first derivatives change direction from monotone increasing to monotone decreasing. Summarizing

$C$-plots by their change in slope turns out to be easier than summarizing them by their change in shape.

Therefore, to summarize the simulations graphically, we apply principal components analysis to the first differences of the simulated $C$-plots. That is, the 2000 vectors of first differences of the $C$-plots are ranked according to their first principal components scores, and the $C$-plots (not the first differences) that correspond to selected quantiles of the principal components scores are plotted. The $C$-plots themselves are shown because in practice one would look at the $C$-plots, not their first differences. The first differences are used only to order the simulated $C$-plots. We denote the $C$-plot corresponding to the $q^{th}$ smallest principal component score by $C_q$, and show the range in $C$-plot behavior by plotting selected $C_q$'s. The idea of using principal components scores to choose which of a large set of curves to plot comes from Jones and Rice (1992).

Because the $C$-plots are ordered according to their first differences, simulated $C$-plots that correspond to small $q$ tend to have a high left tail. As $q$ increases, the left tail flattens out and then the right tail rises (see Figure 1, for example). Convex curves tend to occur for both large and small scores, while moderate scores are typically associated with nonconvex curves. This orderly progression in the shape of the $C_q$'s when ordered by the principal components scores of their first differences simplifies the classification of the $C$-plots considerably, as Figures 3 through 6 show.

To ascertain the fraction of $C$-plots that look convex, the simulated $C$-plots were arranged in $C_q$ order and visually examined. A simulated $C$-plot that was neither obviously convex nor obviously non-convex was deemed convex only if the corresponding $C''(\mu)$ was positive for all but the extreme $\mu$. Since the $C$-plot is intended as a graphical diagnostic rather than as a formal test, this informal assessment is sufficient to show its usefulness. We would not go so far as to assign standard errors on the power of the $C$-plot in the tables below, however.

### 3.3.1 Simulation Results for Poisson Regression

The coefficients of the log-linear Poisson model were taken to be $(\alpha, \beta) = (2, 1)$, giving true means of 2.7, 7.4 and 20.1 at $x = -1, 0, 1$ respectively. The $C$-plots would be better behaved for a smaller range of $\mu_i$, because the uncensored, central regions of the $C_i$'s would overlap more. The $C$-plots would also be better behaved for larger $\mu_i$, because the overdispersion would be more noticeable. The tails of the $C_i$'s were linearized beyond $\hat{\mu}_i \pm 1.5\sqrt{\hat{\mu}_i}$, and the $C$-plots were evaluated at $\mu = \{5, 5.5, \ldots, 18\}$.

Figure 3 shows several of the simulated $C$-plots. In the first panel of Figure 3, $C_{.01}$ is convex, but as $q$ increases the left tails of the $C_q$'s flatten. In the second panel, which has a smaller vertical range, as $q$ increases the right tails flatten and begin to bend upwards, until in the third panel $C_{.96}$ becomes convex. Overall, by examining all the $C$-plots in $C_q$ order, we judged about 6% of the $C$-plots for the pure Poisson simulation to be convex.

Figure 3 about here.

For comparison, Table 1 gives the simulated type I errors for some nominal .05-level score tests for overdispersion. The score tests for a DEF model with a variance inflation factor depending on the mean or on $x$ are given by Smyth (1989) and Ganio and Schafer (1992). The simulated levels of all the tests are close to the nominal .05 level.

Table 1: Simulated power for log-linear Poisson and overdispersed Poisson regression based on 2000 trials. The score tests $S_\alpha$ for random $\alpha$, $S_\beta$ for random $\beta$, $S_{DEF(\mu)}$ for a DEF model with variance depending on $\mu$, and $S_{DEF(x)}$ for a DEF model with variance depending on $x$ all have nominal asymptotic size .05. The entry for the $C$-plot is the fraction of simulated curves judged to be convex. The parameters of the models are given in the text. The estimated standard error of a simulated score test power of $\hat{\rho}$ is $(\hat{\rho}(1-\hat{\rho})/2000)^{1/2}$.

| | $S_\alpha$ | $S_\beta$ | $S_{DEF(\mu)}$ | $S_{DEF(x)}$ | $C$-plot |
|---|---|---|---|---|---|
| Poisson regression | .05 | .03 | .04 | .05 | .06 |
| random intercept | .53 | .32 | .26 | .26 | .53 |
| contamination | .06 | .17 | .19 | .30 | .48 |
| random slope | .52 | .53 | .24 | .18 | .50 |

Figure 4 about here.

Next, consider a Poisson regression with a random normal$(2, \sigma^2 = .05)$ intercept and slope 1. The mean of this model is only 3% larger than it would be if the intercept were fixed at 2, but the variance multiplier ranges from 1.17 to 2.08 as $x$ moves from -1 to 1. Representative $C$-plots are shown in Figure 4. As in Figure 3, as $q$ increases, $C_q$ flattens, then the right tail increases and eventually $C_q$ becomes convex. Only the $C_q$'s in the center panel are not convex. (This is confirmed by the $C''(\mu)$ curves, which are not shown here.) Overall, about 53% of the $C$-plots are convex, which compares favorably with the most powerful score test for the random intercept model. Note that the score test for a random slope also rejects the Poisson model often, even though the slope is fixed in this model. That is, the score tests for random coefficients do not reliably show which coefficient is random. This is not surprising, because the terms in the score test for a random $\beta$ are just a weighted average of the terms in the score test for random $\alpha$.

Now suppose 48 responses follow a Poisson regression with $(\alpha, \beta) = (2, 1)$ and two follow a Poisson regression with $(\alpha, \beta) = (3, 1)$. This contaminated Poisson is close to a random intercept Poisson, but the score test for a random intercept does not reliably detect the overdispersion (see Table 1). The score test for a random slope is more powerful, a DEF score test is more powerful still, and the $C$-plot is most powerful (also see Figure 5). This ordering is not surprising, even though the score test for a random intercept is locally most powerful (Bartoo and Puri 1967). The score test is sensitive to symmetric mixtures with small variances, while the contaminated Poisson model resembles an asymmetric mixture with large variance. Hence, the score test can lack power for a contamination alternative.

Figure 5 about here.

Finally, Table 1 gives the simulated power for a random slope Poisson regression for which $\alpha = 2$ and $\beta$ assumes the values $(.2, 1.2)$ with probabilities $(.2, .8)$. Here, the variance multipliers at $x = -1, 0, 1$ are 1, 1.32 and 2.98 respectively. The $C$-plots and the score tests for random $\alpha$ and random $\beta$ all have nearly the same power, while the DEF score tests are less powerful.

### 3.3.2 Simulation Results for Logistic Regression

The simulated logistic regression model has $\alpha = -1$, $\beta = 2.5$ and $m_1 = \ldots = m_{50} = 6$, so $\mu_i$ ranges from .03 to .82. Again, the $C$-plots would be better behaved for a smaller range of $\mu_i$'s or for larger $m_i$'s. In each run, the tails of the $C_i$'s were linearized beyond $\hat{\mu}_i \pm (\hat{\mu}_i(1-\hat{\mu}_i))^{1/2}$ and $C(\mu)$ was evaluated on $\{.1, .15, \ldots, .8\}$.

Table 2: Simulated power for logistic and overdispersed logistic regression based on 2000 trials. All the score tests have nominal, asymptotic size .05. The parameters of the models are given in the text. The estimated standard error of a simulated score test power of $\hat{\rho}$ is $(\hat{\rho}(1-\hat{\rho})/2000)^{1/2}$.

| | $S_\alpha$ | $S_\beta$ | $S_{DEF(\mu)}$ | $S_{DEF(x)}$ | $C$-plot |
|---|---|---|---|---|---|
| logistic regression | .04 | .03 | .03 | .03 | .06 |
| random intercept | .72 | .49 | .27 | .28 | .70 |
| random slope | .50 | .60 | .27 | .22 | .45 |
| clustered logistic | .82 | .37 | .06 | .07 | .70 |

Table 2 shows the simulated power of several score tests and the $C$-plots for the logistic regression and some overdispersed variants. In the random intercept model, $\alpha$ assumed the values $(-3.0, 2.5)$ with probabilities $(.2, .8)$ and $\beta = 2.5$. In the random slope model, $\alpha = -1$ and $\beta$ assumed the values $(1, 4)$ with probabilities $(.5, .5)$. The $C$-plot detected the random intercept model almost as well as the score test for a random intercept, but the random slope score test detected the random slope more often than the $C$-plot.

Figure 6 about here.

Finally, consider a clustered logistic model in which each $Y_i$ is the total number of positive responses in three clusters of size two each (so $m_i = 6$) and the mean of each cluster at $x_i$

is random with a Beta($.25p(x_i), .25(1-p(x_i))$) distribution where $logit(p(x)) = -1 + 2.5x$. This model has no bias and a constant variance multiplier of 1.8 with respect to the logistic regression model with $\alpha = -1, \beta = 2.5$. Figure 6 shows several $C_q$'s; Table 2 shows the simulated powers of the score tests and $C$-plots. The DEF score tests do not detect the overdispersion, but the $C$-plot and score test for a random intercept and, to a lesser extent, the score test for a random slope do detect the overdispersion. Since the overdispersion is constant in this cluster model, it is not surprising that the DEF score tests cannot detect it. Note that the score test for random $\alpha$ points to a more complicated model of overdispersion than is warranted. $C$-plots, of course, do not attempt to distinguish the source of the overdispersion.

# 4 RELATIVE VARIANCE

## 4.1 Relative Variance Plots

Relative variance curves and tests are based on Theorem 4 below, which gives an approximate mean and variance of $Y$ under a random coefficient GLM. Theorem 4 is proved only for canonical links, but it holds for any increasing link. Recall that each covariate is scaled to lie between -1 and 1.

**Theorem 4** *Suppose $Y$ is a response from a random coefficient GLM with exponential family density $f(\cdot;\eta)$ and at most $\alpha$ and one element of $\beta$ are random. Let $\alpha_0, \beta_0$ and $\sigma^2_\alpha, \sigma^2_\beta, \sigma_{\alpha,\beta}$ be the first and second moments of $\alpha, \beta$, let $\eta_0 = \alpha_0 + x^t\beta_0$, let $\mu_0$ be the mean of $f(\cdot;\eta_0)$, and let $\kappa_j$ be the $j^{th}$ cumulant of $f(\cdot;\eta_0)$. Suppose that for some $\tau$*

$$E[(\alpha-\alpha_0)^r(\beta-\beta_0)^s] = \begin{cases} O(\tau) & \text{if } r+s \le 2 \\ o(\tau) & \text{if } r+s > 2 \end{cases}$$

*and that $\kappa_{j,0} = O(V(\mu_0))$ for $j = 3, 4, 5$. Then, up to terms of order $o(\tau V(\mu_0))$,*

$$\begin{aligned} E[Y] &\approx m\mu_0 + \kappa_{3,0}(\sigma^2_\alpha + 2x\sigma_{\alpha\beta} + x^2\sigma^2_\beta)/2 \\ var[Y] &\approx V(\mu_0) + (\sigma^2_\alpha + 2x\sigma_{\alpha\beta} + x^2\sigma^2_\beta)\left(V^2(\mu_0) + k_{4,0}/2\right). \end{aligned}$$

**Remark 5** *Random coefficient logistic regression and log-linear Poisson regression satisfy the conditions of Theorem 4. For the binomial, all cumulants are $O(m)$, and for the Poisson, $\kappa_j = \mu$ for all $j$. Thus, random coefficient logistic and log-linear Poisson regressions are biased and overdispersed relative to pure logistic and log-linear Poisson regressions, and the bias and overdispersion depend on $x$ even if only $\alpha$ is random. For random coefficient logistic regression with not too much mixing*

$$\begin{aligned} E[Y] - m\mu_0 &\approx .5V(\mu_0)(1-2\mu_0)(\sigma^2_\alpha + 2x\sigma_{\alpha\beta} + x^2\sigma^2_\beta) \\ \frac{1}{V(\mu_0)}\left(\frac{var[Y]}{V(\mu_0)} - 1\right) &\approx (\sigma^2_\alpha + 2x\sigma_{\alpha\beta} + x^2\sigma^2_\beta)\left(1 - \frac{3}{m} + \frac{1}{2V(\mu_0)}\right). \end{aligned}$$

*For random coefficient Poisson regression with not too much mixing,*

$$\begin{aligned} E[Y] - \mu_0 &\approx .5\mu_0(\sigma_\alpha^2 + 2x\sigma_{\alpha\beta} + x^2\sigma_\beta^2) \\ \frac{1}{V(\mu_0)}\left(\frac{var[Y]}{V(\mu_0)} - 1\right) &\approx (\sigma_\alpha^2 + 2x\sigma_{\alpha\beta} + x^2\sigma_\beta^2)\left(1 + \frac{1}{2V(\mu_0)}\right). \end{aligned}$$

Consider the *variance ratio*

$$\mathcal{R}(x, m) = \frac{1}{V(\mu)}\left(\frac{var[Y]}{V(\mu)} - 1\right), \tag{2}$$

where $var[Y]$ is the true variance of $Y$ at $x$ and $V(\mu)$ is the variance of $Y$ at $x$ under the assumed GLM. Theorem 4 shows that $\mathcal{R}(x, m)$ is nearly a positive constant (${\sigma^2}_\alpha$) when only the intercept of the GLM is random and nearly a positive quadratic in $x$ (${\sigma^2}_\beta x^2$) when only the slope of $x$ is random, at least if $V(\mu)$ is large. Our *relative variance curves* depend on these approximations.

Relative variance curves are computed as follows.

1. Fit the standard GLM to $(X_1, Y_1), \ldots, (X_n, Y_n)$ and compute the estimated variances $V(\hat{\mu}_i)$ and the *ordinary residuals* $r_i = y_i - m_i\hat{\mu}_i$ from the fitted GLM mean. The GLM estimates $(\hat{\alpha}, \hat{\beta})$ may be biased estimates of the means of the random coefficients, but the bias can be reduced by fitting the GLM over a restricted range of $x$. In the examples, we fit the GLM with the middle 75% of the $x$ values, for example.

2. Suppose that each covariate and $m$ affect $\mathcal{R}(x, m)$ separately. For example, if there is only one covariate and $m$ varies, assume that

$$\mathcal{R}(x, m) \approx a + s_0(m) + s_1(x), \tag{3}$$

for a constant $a$ and smooth functions $s_0$ and $s_1$ that are centered at zero. This kind of additive model is discussed extensively by Hastie and Tibshirani (1990). If $m$ is constant, $s_0$ is identically zero and, therefore, omitted. If there is only one covariate and $m$ is constant, then $\mathcal{R}(x, m)$ is just a smooth function of $x$.

3. Calculate *estimated variance ratios*

$$\mathcal{R}_i = \frac{r_i^2 - V(\hat{\mu}_i)}{V^2(\hat{\mu}_i)}. \tag{4}$$

Assume equation (3) holds and estimate $a$, $s_0(m)$, and $s_1(x)$ by smoothing $\mathcal{R}_i$ against $(x_i, m_i)$ using weights $w_i = 1/var(\mathcal{R}_i)$. Under a log-linear Poisson model $var(\mathcal{R}_i) \approx (1 + 2\hat{\mu}_i)\hat{\mu}_i^{-3}$; under a logistic regression model $var(\mathcal{R}_i) \approx (1 + 2m_i\hat{\mu}_i(1 - \hat{\mu}_i)(1 - 3/m_i))(m_i\hat{\mu}_i(1 - \hat{\mu}_i))^{-3}$. The smoothing can be done by the function *gam* in the statistical language *S* (Chambers and Hastie, 1991), for instance.

4. The *relative variance curves* $\hat{\mathcal{R}}_0(m)$ and $\hat{\mathcal{R}}_1(x)$ are the estimated effects

$$\begin{aligned}\hat{\mathcal{R}}_0(m) &= \hat{a} + \hat{s}_0(m) \\ \hat{\mathcal{R}}_1(x) &= \hat{a} + \hat{s}_1(x).\end{aligned}$$

Sometimes, the adjusted residuals $\mathcal{R}_i - \hat{\mathcal{R}}_1(x)$ and $\mathcal{R}_i - \hat{\mathcal{R}}_0(m)$ are also helpful.

Figure 2 shows $\hat{\mathcal{R}}_0(m)$ and $\hat{\mathcal{R}}_1(x)$ for the toxoplasmosis data discussed in Section 2; some simulated relative variance curves are shown in Section 4.3 below.

## 4.2 Relative Variance Tests

Hypotheses about overdispersion can be translated into hypotheses about the variance ratio $\mathcal{R}(x, m)$, at least approximately. Under a pure GLM, $H_0 : \mathcal{R}(x, m) = 0$ holds. When the intercept of the GLM is random, $H_1 : \mathcal{R}(x, m) = c_1 > 0$ holds approximately, and when the coefficient of $x$ is random $H_2 : \mathcal{R}(x, m) = c_2\, x^2$ for some $c_2 > 0$ holds approximately (see Theorem 4). For overdispersion that cannot be attributed to a random intercept or slope, appropriate hypotheses are $H_0^* : \mathcal{R}(x, m) = c_1 + c_2 x^2$ where $c_i \geq 0$ and $H_3 : \mathcal{R}(x, m) = a + s_0(m) + s_1(x)$ where $s_1(x)$ is not a positive quadratic in $x$.

Since the $\mathcal{R}_i$'s (eqn. (4)) are simple to compute and approximately $N(\mathcal{R}(x_i, m_i), 1/w_i)$, they are a natural foundation for testing the hypotheses of interest. (Step 3 above gives $w_i$.) First, fit the model $E[\mathcal{R}_i] = c_1 + c_2 x^2$ to $(x_i, \mathcal{R}_i)$ using weights $w_i$ and compute the standard, one-sided linear model $t$-statistics for $c_1 = 0$ and $c_2 = 0$. Then, test $c_1 = 0$ against $c_1 > 0$ assuming that $c_2 \neq 0$ and test $c_2 = 0$ against $c_2 > 0$ assuming that $c_1 \neq 0$ using the residual standard deviation with $n - 2$ degrees of freedom from the full model in both $t$-statistics. The alternative $H_3$ can be tested against $H_0^*$ with the approximate $F$-statistic

$$F = \frac{[RSS(H_0^*) - RSS(H_3)]/p}{RSS(H_3)/(n - p - 2)}$$

where $RSS(H)$ is the residual sum of squares (with weights $w_i$) associated with the hypothesis $H$ and $p$ is the number of degrees of freedom used to fit the smooth to $\mathcal{R}_i$ after fitting a nonnegative quadratic. The $F$-test statistic is available with the *gam* function in $S$, for example. If including the smooth term drops the residual sum of squares significantly, then $H_0^*$ is rejected in favor of $H_3$.

In principle, these tests can be extended to multiple covariates by fitting a fuller model for $E[\mathcal{R}_i]$ and testing for the significance of several $c_j$'s. If the covariates $x$ and $z$ are independent and the slope for $z$ is random, then $\hat{\mathcal{R}}_1(x)$ should look like a positive constant while the relative variance curve for $z$ should look quadratic. Of course, distinguishing a random slope on $x$ from a random slope on $z$ is nearly impossible if $x$ and $z$ are strongly correlated.

## 4.3 Relative Variance Simulations

As in Section 3.3, several simulations with 2000 runs each were carried out in $S$. In each run, there was one covariate $x$ with values on a uniform, fixed grid between -1 and 1. Log-linear Poisson and random coefficient Poisson simulations are discussed in Section 4.3.1; logistic and random coefficient logistic regression simulations in Section 4.3.2. The binomial size parameters $m_i$ were allowed to vary in the logistic simulations, so both $\hat{\mathcal{R}}_1(x)$ and $\hat{\mathcal{R}}_0(m)$ are considered in Section 4.3.2.

Since the objective is to identify the source of the overdispersion rather than just to note the presence of overdispersion, the performance of the relative variance tests is summarized by the fraction of runs that correctly identify which coefficient (if any) is random. For example, under a random intercept GLM the score test for a random intercept is evaluated by the fraction of runs in which it rejects and the score test for a random slope does not reject. Note that the power of a test for a random coefficient can exceed its ability to identify the random coefficient correctly, because the test is penalized when only the intercept is random but the tests for random intercept and random slope both reject, for example.

The simulated relative variance curves are summarized graphically. Selected $\hat{\mathcal{R}}_1(x)$ curves are found by applying principal components analysis to the 2000 $\hat{\mathcal{R}}_1(x)$ curves and the $\hat{\mathcal{R}}_1(x)$ that correspond to the $\{.1, .25, .5, .75, .9\}$-quantiles of the principal components scores are plotted. Selected $\hat{\mathcal{R}}_0(m)$ curves for logistic regression are also found by using principal components analysis. Because $\hat{\mathcal{R}}_1(x)$ and $\hat{\mathcal{R}}_0(m)$ should be approximately constant when all coefficients are fixed or when only the intercept is random, the ordered $\hat{\mathcal{R}}_1(x)$ and $\hat{\mathcal{R}}_0(m)$ curves do not necessarily change shape in an orderly manner as $q$ increases. Nonetheless, the ordered $\hat{\mathcal{R}}_1(x)$ and $\hat{\mathcal{R}}_0(m)$ curves shown represent a range of likely behaviors.

### 4.3.1 Results of the Poisson Simulations

The log-linear Poisson regression has $(\alpha, \beta) = (2, 1)$ and $n = 50$ responses. In the random intercept model, $\beta = 1$ and $\alpha \sim N(2, \sigma_\alpha^2)$ where $\sigma_\alpha^2$ is either .05, .1 or .25. In the random slope model, $\alpha = 2$ and $\beta \sim N(1, \sigma_\beta^2)$ where $\sigma_\beta^2$ is either .25, .5 or 1.

Table 3 summarizes the results. For less variable random intercepts, the locally most powerful score test and the relative variance test are roughly equivalent. As $\sigma_\alpha^2$ increases, however, the relative variance test more often finds a random intercept but not a random slope while the score tests cannot distinguish the random intercept from a random slope. In this sense, the relative variance tests outperform the score tests.

The relative variance test also detects random slopes better than the score test does. Random slopes are harder to detect than random intercepts, though, and the relative variance test does not continue to improve as $\sigma_\beta^2$ increases. Moreover, the relative variance tests often suggested $\mathcal{R}(x, m)$ should be nonparametric when $\sigma_\beta^2$ was positive, implying (wrongly) that the overdispersion should be modeled as something other than a random intercept or slope. The simulated $\hat{\mathcal{R}}_1(x)$ curves often look roughly quadratic, however, even if the $R_\beta$ does not often reject, suggesting that replacing the relative variance curve with a relative variance test is not always a good idea. (See Figure 7, for example.)

Table 3: Simulated fraction of runs in which a random coefficient was identified for a loglinear Poisson model. $S_\alpha$ and $S_\beta$ refer to score tests for random coefficient Poissons; $R_\alpha$ and $R_\beta$ refer to relative variance tests for random coefficient Poissons. The column labeled $R_{np}$ shows the simulated power of the relative variance test for $H_3$. All tests have nominal size .05, and the standard error of a reported value of $\rho$ is $(\rho(1-\rho)/2000)^{1/2}$ for all tests.

| | $S_\alpha$ | $S_\beta$ | $R_\alpha$ | $R_\beta$ | $R_{np}$ |
|---|---|---|---|---|---|
| Poisson(2,1) | .07 | .03 | .04 | .10 | .10 |
| $\sigma_\alpha^2 = .05$ | .26 | .03 | .29 | .09 | .13 |
| $\sigma_\alpha^2 = .1$ | .26 | .01 | .54 | .09 | .12 |
| $\sigma_\alpha^2 = .25$ | .04 | .00 | .80 | .04 | .11 |
| $\sigma_\beta^2 = .25$ | .06 | .02 | .14 | .36 | .31 |
| $\sigma_\beta^2 = .5$ | .02 | .00 | .19 | .40 | .33 |
| $\sigma_\beta^2 = 1$ | .00 | .00 | .25 | .35 | .37 |

Figure 7 shows selected $\hat{\mathcal{R}}_1(x)$ curves for these simulations. In the random intercept model, $\sigma_\alpha^2 = .1$. In the random slope model, $\sigma_\beta^2 = .5$. As would be hoped, the $\hat{\mathcal{R}}_1(x)$'s from the pure Poisson and random intercept Poisson vary around the ideal constants 0 and .1 respectively. The average $n^{-1}\sum \hat{\mathcal{R}}_1(x_i)$ is -.004 (with a simulated standard error of .004) for the Poisson and .096 (with a simulated standard error of .04) for the random intercept Poisson. For the random slope Poisson, three of the five $\hat{\mathcal{R}}_1(x)$ curves shown are roughly quadratic, as Theorem 4 promises, but only one of the corresponding relative variance tests has a p-value less than .1. Thus, the relative variance tests may miss information provided by the relative variance curves.

Figure 7 about here.

### 4.3.2 Results of the Logistic Simulations

For each simulation, $n = 50$ or 100 and $m_i$ varies with $i$. The covariate values $(x_1, \ldots, x_n)$ were not changed from run-to-run, but the $m_i$'s were randomly permuted from run-to-run. For $n = 50$, $(m_1, \ldots, m_n)$ is a random permutation of $(4, 5, 6, 7, 6, 7, 6, 5, 4)$ copies of $(6, 7, \cdots, 14)$ respectively; for $n = 100$ there were twice as many copies of $(6, 7, \cdots, 14)$. For the pure logistic regression, $n = 50$ and $(\alpha, \beta) = (0, 2.5)$. For the random intercept logistic regression, $n = 50$, $\beta = 2.5$ and $\alpha \sim N(0, \sigma_\alpha^2)$ where $\sigma_\alpha^2$ is either .75, 1 or 1.5. For the random slope logistic regression, $n = 50$ or 100, $\alpha = 0$ and $\beta \sim N(0, \sigma_\beta^2)$ where $\sigma_\alpha^2$ is either 1, 1.5 or 2. Table 4 and Figures 8 and 9 summarize the results.

Table 4 shows that the relative variance test's ability to distinguish a random intercept from a random slope improves for larger $\sigma_\alpha^2$ but the ability of the locally most powerful score test degrades. Moreover, the test for a nonparametric $\mathcal{R}(x,m)$ seldom rejects and seldom (wrongly) suggests that there is overdispersion that cannot be attributed to a random coefficient. Under the logistic GLM, $\hat{\mathcal{R}}_0(m)$ and $\hat{\mathcal{R}}_1(x)$ have means -.020 and -.012 respectively (with simulated standard errors of .10 and .15 respectively), which seems satisfactory since the ideal value is 0. For a random intercept model with $\sigma_\alpha^2 = 1$, the $\hat{\mathcal{R}}_1(x)$ curves vary around an average level of .703 (with a simulated standard error of .23), which is lower than the ideal value of 1.

Table 4: Simulated fraction of runs in which a random coefficient was distinguished for pure and random coefficient logistic regression models. For a random intercept model, for example, high values for the $\alpha$ tests and low values for the $\beta$ tests are desirable. All tests have nominal size .05. The column labeled $R_{np}$ shows the simulated power of the relative variance test for $H_3$. All tests have nominal size .05, and the standard error of a reported value of $\rho$ is $(\hat{\rho}(1-\hat{\rho}))^{1/2}$ for all tests.

| | $S_\alpha$ | $S_\beta$ | $R_\alpha$ | $R_\beta$ | $R_{np}$ |
|---|---|---|---|---|---|
| | | | $n = 50$ | | |
| logistic(0,2.5) | .07 | .04 | .04 | .04 | .07 |
| $\sigma_\alpha^2 = .75$ | .51 | .01 | .85 | .10 | .09 |
| $\sigma_\alpha^2 = 1$ | .29 | .00 | .88 | .01 | .09 |
| $\sigma_\alpha^2 = 1.5$ | .03 | .00 | .88 | .00 | .09 |
| $\sigma_\beta^2 = 1$ | .11 | .20 | .10 | .05 | .17 |
| $\sigma_\beta^2 = 1.5$ | .07 | .09 | .23 | .04 | .30 |
| $\sigma_\beta^2 = 2$ | .03 | .02 | .45 | .03 | .43 |
| | | | $n = 100$ | | |
| $\sigma_\beta^2 = 1$ | .09 | .21 | .20 | .05 | .30 |
| $\sigma_\beta^2 = 1.5$ | .02 | .02 | .51 | .03 | .60 |
| $\sigma_\beta^2 = 2$ | .00 | .00 | .75 | .01 | .79 |

Unfortunately, Table 4 also shows that the relative variance tests are unable to detect a random slope, even if the sample size is doubled from 50 to 100. Instead, the relative variance tests suggest that $\mathcal{R}(x,m)$ is either a positive constant or nonparametric. That the $R_\alpha$ test rejects so often is not surprising, because the $R_\beta$ test has low power and $\hat{\mathcal{R}}_1(x)$ tends to lie above zero.

Figure 8 about here.

Figure 9, which summarizes the random slope logistic regression with $\sigma^2 = 1.5$, shows that plotting the relative variance *curves* can be a better way to detect a random slope than computing the relative variance *tests*. Although two of the curves in the plot for $n = 50$ look roughly quadratic, the smallest corresponding relative variance $R_\beta$ p-value is .43. In the $n = 100$ plot, four of the curves look quadratic, but the smallest $R_\beta$ p-value corresponding to these five curves is .19. That is, none of the corresponding relative variance tests would reject, despite the strong evidence in the relative variance curves. On the other hand, the largest random intercept or random slope score test p-value for the curves shown is .07. Hence, the score tests corresponding to the relative variance curves shown in Figure 9 are unable to distinguish random $\beta$ from a random $\alpha$, even though many of the relative variance curves shown strongly suggest a random slope rather than a random intercept.

Figure 9 about here.

It is not surprising that the relative variance tests and curves behave better in the Poisson simulation than in the logistic simulation. In the logistic simulation, there are two variables $x_i$ and $m_i$ to consider, in the Poisson only $x_i$. Moreover, the additive assumption (3) is not needed in the Poisson simulation, but it is needed in the logistic regression.

## 5 DISCUSSION

This paper introduces new tools for detecting and understanding overdispersion in GLMs: $C$-plots, relative variance curves and relative variance tests. The $C$-plot can be calculated easily for any GLM, without making any assumption about the source of overdispersion or pattern of overdispersion. Yet $C$-plots are sensitive to many kinds of overdispersion, even if the variance inflation factor is constant. Simulations show that $C$-plots can be nearly as sensitive as the locally most powerful score tests for random coefficients under conditions that are nearly ideal for the score tests, and more sensitive than the score tests under other conditions. The $C$-plot cannot distinguish different sources of overdispersion, but score tests often cannot do that either. We note that the $C$-plot has been derived using properties of GLMs rather than by extending familiar diagnostics for linear models to GLMs. We like the fact that the $C$-plot is defined explicitly for GLMs, but it is true that the $C$-plot does not reduce to a recognizable diagnostic for gaussian linear models, which do not have a specified mean–variance relationship.

Relative variance tests can also be computed easily for any GLM, without fitting an overdispersed GLM. All that is required is to fit the GLM, find the ordinary residuals from the mean, compute estimated variance ratios, and compute weighted t-statistics for a simple linear model. Computing relative variance curves requires smoothing the relative variance estimates, but smoothing is often no more difficult than fitting the GLM. The simulations show that the extra computing needed to compute smooth relative variance curves is worthwhile, because the relative variance curves may detect random slopes better than the tests do.

Relative variance curves and tests have been designed for detecting random coefficients, but they could also give information about other sources of overdispersion. For example, in a log-linear Poisson quasi-likelihood model with constant variance multiplier $\gamma$, $\mathcal{R}(x,m) = (\gamma-1)e^{-(\alpha+x^t\beta)}$. Thus, if $\hat{\mathcal{R}}_1(x)$ is approximately an exponential in $x$, we would choose the simplest model of overdispersion and use quasi-likelihood.

On the other hand, neither the relative variance curves nor the relative variance tests can distinguish two overdispersed GLMs with the same variance pattern. For example, they cannot distinguish beta-binomial or double logistic GLMs with the same variance functions. This may be unimportant, however, since the goal may be to find an adequate model rather than to pinpoint the "true" underlying model. Similarly, because our tools focus only on overdispersion, they cannot distinguish poor fits caused by overdispersion and poor fits caused by inappropriate links. In practice, since it is often easier to fit GLMs with different links than to fit overdispersed GLMs, we would first find the best link and then move on to an overdispersed GLM only if there was substantial overdispersion from the GLM with the best link.

In short, despite their inability to resolve all issues about overdispersion, we find $C$-plots and relative variance tests and curves to be simple tools that can be routinely used without much additional effort to obtain some understanding of whether overdispersion is present and, if so, how it should be modeled.

## APPENDIX: PROOFS

**Proof of Theorem 2**

To simplify the notation, assume the GLM has its canonical link $l$. If not, the natural parameter of $f$ must be written as $l(h^{-1}(\alpha + x^t\beta))$ rather than as $\alpha + x^t\beta$.

Because $C(\mu)$ is an average over the joint distribution of $X, Y$, it can be approximated by an integral over $X, Y$ for large $n$. Under a GLM with random coefficients, $Y$ has exponential family density $f$ given $x$ and $(\alpha, \beta)$. If $(\alpha,\beta) \sim Q$, then given $x$ alone $Y$ has density

$$f(y; x, Q) = \int f(y; \alpha + x^t\beta)\, dQ(\alpha,\beta).$$

Letting $F_X$ denote the marginal distribution of $X$

$$\begin{aligned} C(\mu) &\approx \int_{\mathcal{X}}\int_{\mathcal{Y}} \Big[\frac{f(y; l(\mu))}{f(y; \hat{\alpha} + x^t\hat{\beta})}\Big] f(y; x, Q)\, d\gamma(y)\, dF_X(x) \\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} \Big[\frac{f(y; x, Q)}{f(y; \hat{\alpha} + x^t\hat{\beta})}\Big] f(y; l(\mu))\, d\gamma(y)\, dF_X(x). \end{aligned}$$

The last term in square brackets is a convex function of $y$ for each $x$ (Shaked 1980 and Lindsay 1986). Because $f(y; l(\mu))$ is an exponential family density, the inner integral is then a convex function of $\mu$ for each $x$ (Karlin 1968, p. 23). Finally, a nonnegative sum of convex functions is convex, so the integral approximation to $C(\mu)$ is convex.

Under a random mean GLM, $Y$ has density $f$ given $\alpha + x_i\beta$, $\alpha + x_i\beta \sim Q$, and unconditionally $Y$ has density $f(y; Q) = \int f(y; z)\, dQ(z)$. Then

$$\begin{aligned} C(\mu) &\approx \int_{\mathcal{X}} \int_{\mathcal{Y}} \left[\frac{f(y; l(\mu))}{f(y; \hat{\alpha} + x^t\hat{\beta})}\right] f(y; Q)\, d\gamma(y)\, dF_X(x) \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \left[\frac{f(y; Q)}{f(y; \hat{\alpha} + x^t\hat{\beta})}\right] f(y; l(\mu))\, d\gamma(y)\, dF_X(x). \end{aligned}$$

To complete the proof, appeal to the results of Shaked (1980), Lindsay (1986), and Karlin (1986) again.

**Proof of Theorem 3**

An overdispersed DEF density can be expressed as

$$f(y; \eta, \tau) = e^{y\eta + \tau T(y) - \rho(\eta,\tau)}\, d\gamma(y) \tag{5}$$

for some natural parameter $\eta$, constant $\tau$, convex function $T$, and supporting measure $\gamma$ (Gelfand and Dalal 1990). Under a DEF regression model with its canonical link $l$, $Y_1, \ldots, Y_n$ are independent with DEF density $f(y; \eta_i, \tau)$ and $l$ is the canonical link for the exponential family density $f(y; \eta_i)$. As in the proof of Theorem 2, to simplify the notation suppose that the GLM has its canonical link $l$.

Since $T$ in the DEF density (equation (5)) is convex, $f(y; \eta, \tau)/f(y; \eta)$ is a convex function of $y$ for fixed $\eta, \tau$, where $f(y; \eta)$ is the exponential family density with natural parameter $\eta$. Then,

$$\begin{aligned} C(\mu) &\approx \int_{\mathcal{X}} \int_{\mathcal{Y}} \left[\frac{f(y; l(\mu))}{f(y; \hat{\alpha} + x^t\hat{\beta})}\right] f(y; \alpha + x^t\beta), \tau)\, d\gamma(y)\, dF(x) \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \left[\frac{f(y; \alpha + x^t\beta, \tau)}{f(y; \hat{\alpha} + x^t\hat{\beta})}\right] f(y; l(\mu))\, d\gamma(y)\, dF(x) \\ &\approx \int_{\mathcal{X}} \int_{\mathcal{Y}} \left[\frac{f(y; \alpha + x^t\beta, \tau)}{f(y; \alpha + x^t\beta)}\right] f(y; l(\mu))\, d\gamma(y)\, dF(x). \end{aligned}$$

The right side is convex because the last term in square brackets is convex in $y$.

**Proof of Theorem 4**

First, recall that $\kappa_j$ equals the $j^{th}$ derivative $\psi^{(j)}$ of the term $\psi$ in equation (1).

Because the conditional density of $Y$ given $\eta = \alpha + x^t\beta$ is in the exponential family, the unconditional mean and variance of $Y$ are

$$\begin{aligned} E[Y] &= E[\psi'(\eta)] \\ var[Y] &= E[\psi''(\eta)] + var[\psi'(\eta)]. \end{aligned}$$

Expanding $\psi'(\eta)$ and $\psi''(\eta)$ about $\eta_0$ gives

$$\psi'(\eta) = \psi'(\eta_0) + (\eta - \eta_0)\psi''(\eta) + \frac{(\eta - \eta_0)^2}{2}\psi^{(3)}(\eta) + \frac{(\eta - \eta_0)^3}{6}\psi^{(4)}(\eta_1)$$

$$\psi''(\eta) \;=\; \psi''(\eta_0) + (\eta - \eta_0)\psi^{(3)}(\eta) + \frac{(\eta - \eta_0)^2}{2}\psi^{(4)}(\eta) + \frac{(\eta - \eta_0)^3}{6}\psi^{(5)}(\eta_2)$$

for some $\eta_1, \eta_2$ between $\eta$ and $\eta_0$. Therefore, ignoring terms of $o(\tau V(\mu_0))$,

$$\begin{aligned} E[\psi'(\eta)] &\approx \psi'(\eta_0) + .5(\sigma_\alpha^2 + 2x\sigma_{\alpha\beta} + x^2\sigma_\beta^2)\psi^{(3)}(\eta_0) \\ var[\psi'(\eta)] &\approx (\sigma_\alpha^2 + 2x\sigma_{\alpha\beta} + x^2\sigma_\beta^2)(\psi''(\eta_0))^2 \\ E[\psi''(\eta)] &\approx \psi''(\eta_0) + .5(\sigma_\alpha^2 + 2x\sigma_{\alpha\beta} + x^2\sigma_\beta^2)\psi^{(4)}(\eta_0), \end{aligned}$$

which completes the proof.

## References


[1] Anderson, D. and Aitkin, M. (1968), "Variance Component Models with Binary Response: Interviewer Variability," *Journal of the Royal Statistical Society, Series B*, 47, 203-210.

[2] Bartoo, J.B. and Puri,P.S. (1967), "On Optimal Asymptotic Tests of Composite Statistical Hypotheses," *Annals of Mathematical Statistics*, 38, 1845-1852.

[3] Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988), *The New S Language: A Programming Environment for Data Analysis and Graphics*, Pacific Grove, CA: Wadsworth & Brooks.

[4] Breslow, N. E. (1984), "Extra-Poisson Variation in Log-Linear Models," *Applied Statistics*, 33, 38-44.

[5] Butler, S. M. and Louis, T. A. (1992), "Random effects models with non-parametric priors," *Statistics in Medicine*, 11, 1981-2000.

[6] Chambers, J.M. and Hastie, T. (eds.) (1991), *Statistical Models in S*, Pacific Grove, CA: Wadsworth & Brooks/Cole.

[7] Cox, D. R. (1983), "Some Remarks on Overdispersion,"*Biometrika*, 70, 269-274.

[8] Dean, C. B. (1992), "Testing for Overdispersion in Poisson and Binomial Regression Models," *Journal of the American Statistical Association*, 87, 451-457.

[9] Efron, B. (1986), "Double Exponential Families and Their Use in Generalized Linear Regression," *Journal of the American Statistical Association*, 81, 709-721.

[10] Follmann, D. A. and Lambert, D. (1989), "Generalizing Logistic Regression by Nonparametric Mixing," *Journal of the American Statistical Association*, 84, 295-300.

[11] Ganio, L. M. and Schafer, D. W. (1992), "Diagnostics for Overdispersion," *Journal of the American Statistical Association*, 87, 795-804.

[12] Gelfand, A. E. and Dalal, S. R. A. (1990), "A Note on Overdispersed Exponential Families," *Biometrika*, 77, 55-64.

[13] Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, New York: Chapman Hall.

[14] Jones, M. C. and Rice, J. A. (1992), "Displaying the Important Features of a Large Collection of Similar Curves," *The American Statistician*, 46, 140-145.

[15] Karlin, S. (1968), *Total Positivity (Vol. 1)*, Palo Alto, CA: Stanford University Press.

[16] Lindsay, B. G. (1986), "Exponential Family Mixture Models (With Least Squares Estimators)," *Annals of Statistics*, 14, 124-137.

[17] Lindsay, B. G. and Roeder, K. (1992), "Residual Diagnostics for Mixture Models," *Journal of the American Statistical Association*, 87, 785-794.

[18] McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models* (2nd. ed.), New York: Chapman Hall.

[19] Nelder, J. A. and Pregibon, D. (1987), "An Extended Quasi-Likelihood Function," *Biometrika*, 74, 221-232.

[20] Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991), "A Comparison of Cluster-Specific and Population Averaged Approaches for Analyzing Correlated Binary Data," *International Statistical Review*, 59, 22-35.

[21] Pierce, D.A. and Sands, B.R. (1975) "Extra-Binomial Variation in Binary Data," Technical Report 46, Oregon State University, Dept. of Statistics.

[22] Shaked, M. (1980), "On Mixtures From Exponential Families," *Journal of the Royal Statistical Society, Series B*, 42, 192-198.

[23] Smith, P. J. and Heitjin, D. F. (1993), "Testing and Adjusting for Departures from Nominal Dispersion in Generalized Linear Models," *Applied Statistics*, 42, 31-42.

[24] Smyth, G. K. (1989), "Generalized Linear Models with Varying Dispersion," *Journal of the Royal Statistical Society, Series B*,51, 47-60.

[25] Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986), "Optimally bounded score functions for generalized linear models with applications to logistic regression," *Biometrika*, 73, 413-424.

[26] Wedderburn, R. W. M. (1974), "Quasilikelihood Functions, Generalized Linear Models and the Gauss-Newton Method,"*Biometrika*, 61, 439-477.

[27] Zelterman, D. and Chen, C-F. (1988), "Homogeneity Tests Against Central Mixture Alternatives," *Journal of the American Statistical Association*, 83, 179-182.

# FIGURES

Figure 1. Convexity plots for two fits to the toxoplasmosis data. The solid line is the $C$-plot for the linear logistic fit; the dotted line is the $C$-plot for the signed quadratic fit. Because convexity suggests overdispersion, there could be overdispersion from the linear logistic regression, perhaps caused by omitting an important explanatory variable in the specification of the mean. But there is no evidence of overdispersion from the signed quadratic.

Figure 2. Relative variance plots for the linear logistic fit to the toxoplasmosis data. $\hat{R}_1(x)$ is quadratic, suggesting that $x$ has a random slope or an interaction between $x$ and an uncorrelated variable has been omitted. $\hat{R}_0(m)$ is more nearly zero, suggesting that the overdispersion is not as dependent on $m$.

Figure 3. Selected $C_q$'s from the log-linear Poisson simulation. The $q$ for each $C_q$ is given on the right side of the panel. Overall, about 6% of the simulated $C$-plots were judged convex.

Figure 4. Selected $C_q$'s from the random intercept Poisson simulation. In the first panel, the $C_q$'s flatten as $q$ increases; in the second and third panels the right tail of $C_q$ increases as $q$ increases. The $C_q$'s in the first and last plot are convex; those in the middle plot are not.

Figure 5. Selected $C_q$'s from the contaminated Poisson regression model with two outliers. The $C_q$'s for $q > .5$ were judged to be convex.

Figure 6. Selected $C_q$'s from the clustered logistic regression model. The $C_q$'s are convex for $q < .2$ and $q \geq .5$. In the first panel, the left tail of $C_q$ flattens as $q$ increases. There, the $C_q$'s are labeled by the values of their left tails instead of their right tails to better distinguish the curves. In the second panel, the right tail of $C_q$ increases as $q$ increases until the $C_q$'s become convex again.

Figure 7. Selected $\hat{\mathcal{R}}_1(x)$ from the pure Poisson and random coefficient Poisson regressions. The dashed lines denote the $\{.1, .9\}$-$\hat{\mathcal{R}}_1(x)$; the dotted the $\{.25, .75\}$-$\hat{\mathcal{R}}_1(x)$; and the solid the median $\hat{\mathcal{R}}_1(x)$. Ideally, $\hat{\mathcal{R}}_1(x)$ would be a straight line at zero for the Poisson, a straight line at .1 for the random intercept ($\sigma_\alpha^2 = .1$) Poisson, and a quadratic with coefficient .5 for the random slope ($\sigma_\beta^2 = .5$) Poisson. The tails for the most extreme random slope curves continue to look quadratic, but they are not shown to preserve the detail in the less extreme curves.

Figure 8. Selected $\hat{\mathcal{R}}_0(m)$ (first row) and $\hat{\mathcal{R}}_1(x)$ (second row) from the pure logistic (first column) and random intercept logistic (second column) regressions. In the random intercept model, $\sigma_\alpha^2 = 1$. The dashed lines denote the $\{.1, .9\}$-$\hat{\mathcal{R}}_1(x)$; the dotted the $\{.25, .75\}$-$\hat{\mathcal{R}}_1(x)$; and the solid the .5-$\hat{\mathcal{R}}_1(x)$.

Figure 9. Selected $\hat{\mathcal{R}}_1(x)$ from the random slope logistic regressions with $\sigma_\beta^2 = 1.5$. The dashed lines denote the $\{.1, .9\}$-$\hat{\mathcal{R}}_1(x)$; the dotted the $\{.25, .75\}$-$\hat{\mathcal{R}}_1(x)$; and the solid the median $\hat{\mathcal{R}}_1(x)$. Ideally, $\hat{\mathcal{R}}_1(x)$ would be a quadratic with coefficient 1.5. The tails for the most extreme random slope curves continue to look quadratic, but they are not shown to preserve the detail in the less extreme curves.

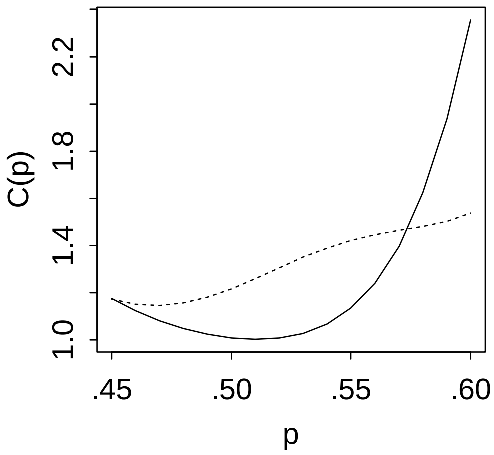



Figure 1:

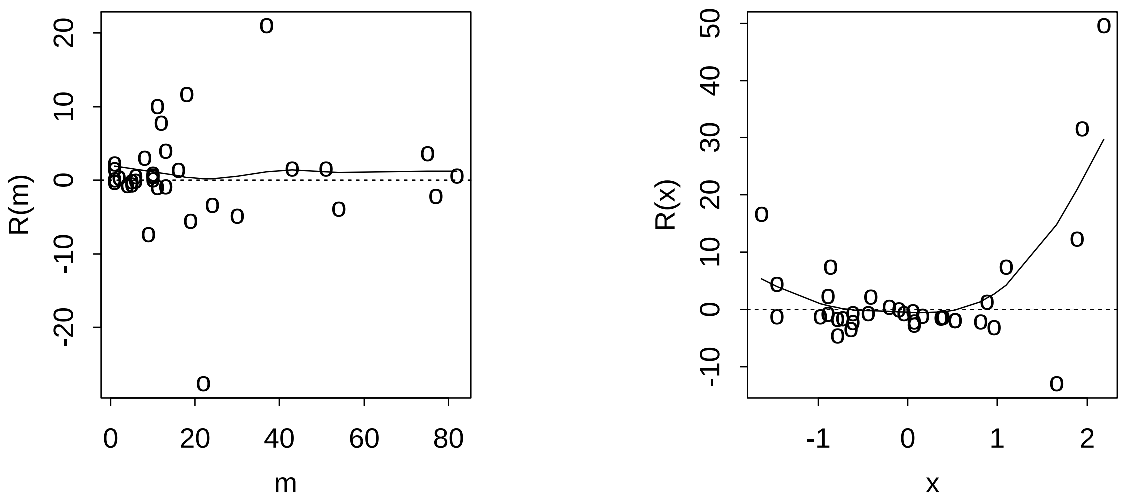



Figure 2:

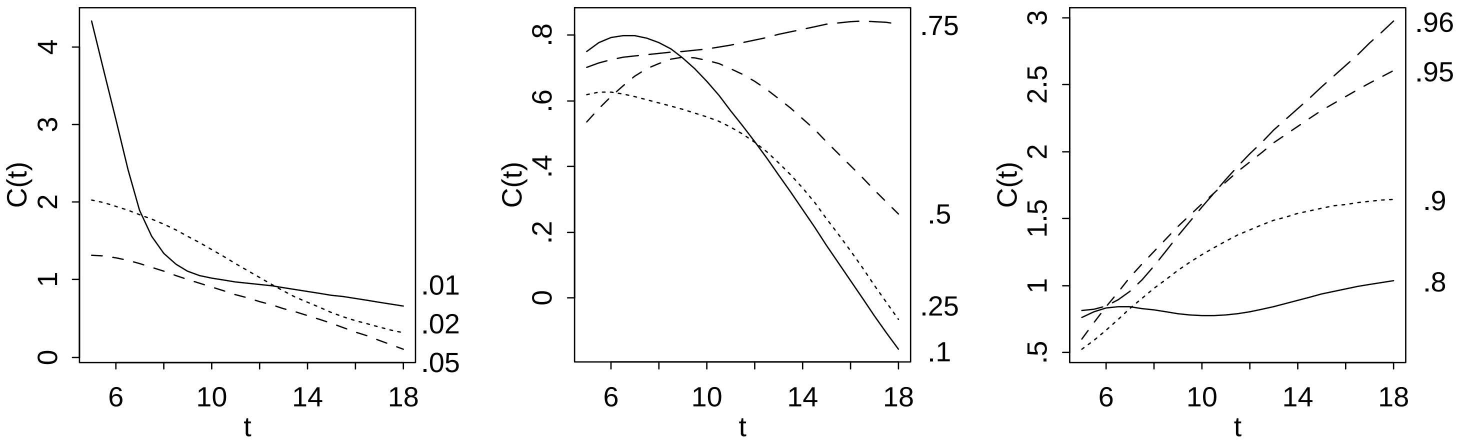



Figure 3:

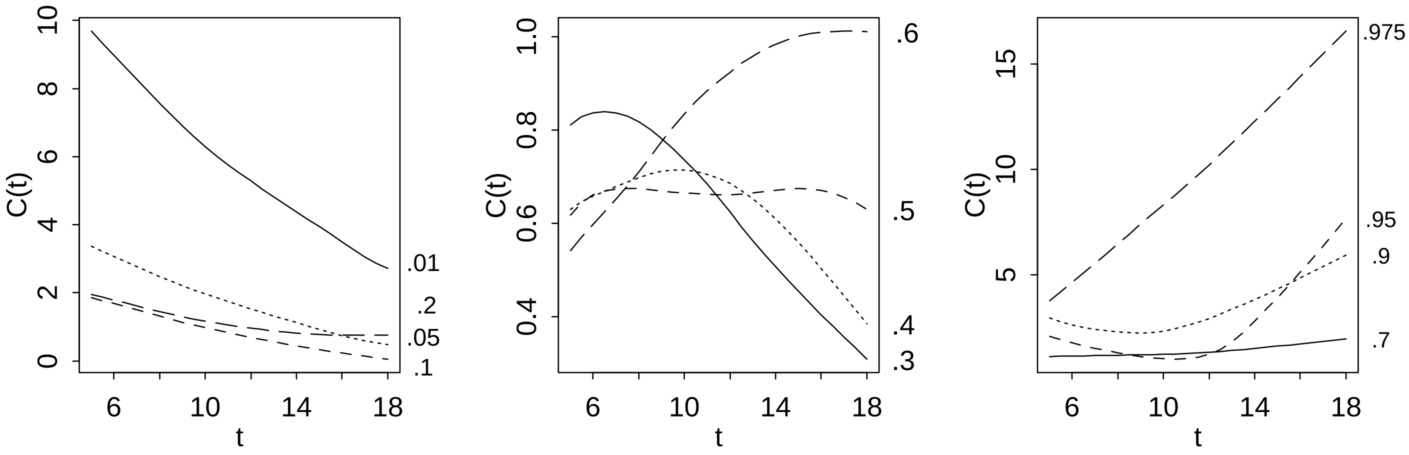



Figure 4:

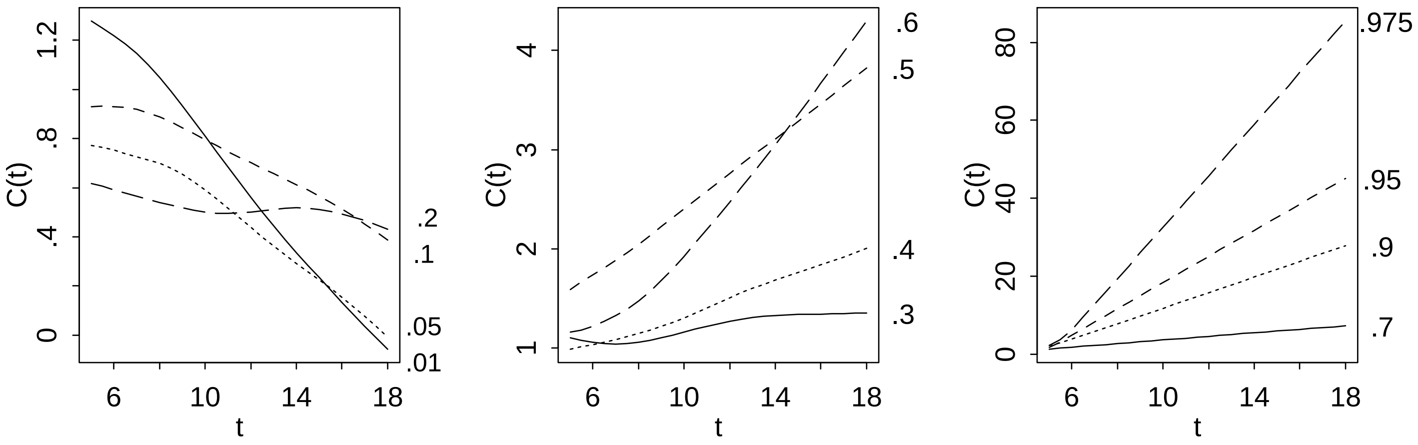
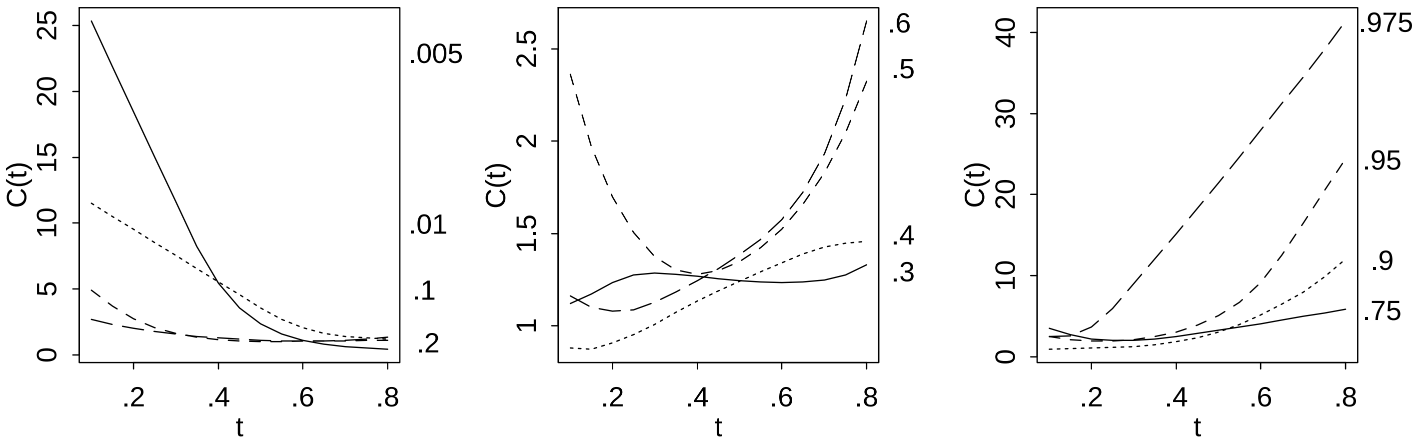



Figure 5:



C(t)
0
5
10
15
20
25
.2
.4
.6
.8
t
.005
.01
.1
.2
C(t)
1
1.5
2
2.5
.2
.4
.6
.8
t
.6
.5
.4
.3
C(t)
0
10
20
30
40
.2
.4
.6
.8
t
.975
.95
.9
.75


Figure 6:

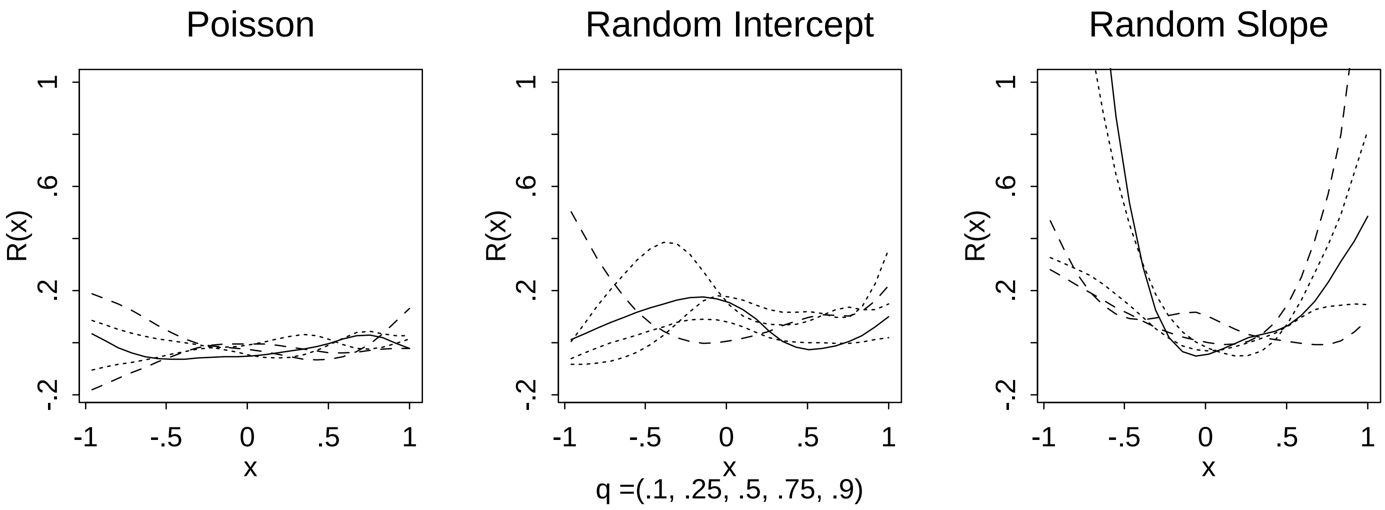

Poisson
Random Intercept
Random Slope
R(x)
1
.6
.2
-.2
-1
-.5
0
.5
x
q =(.1, .25, .5, .75, .9)

Figure 7:

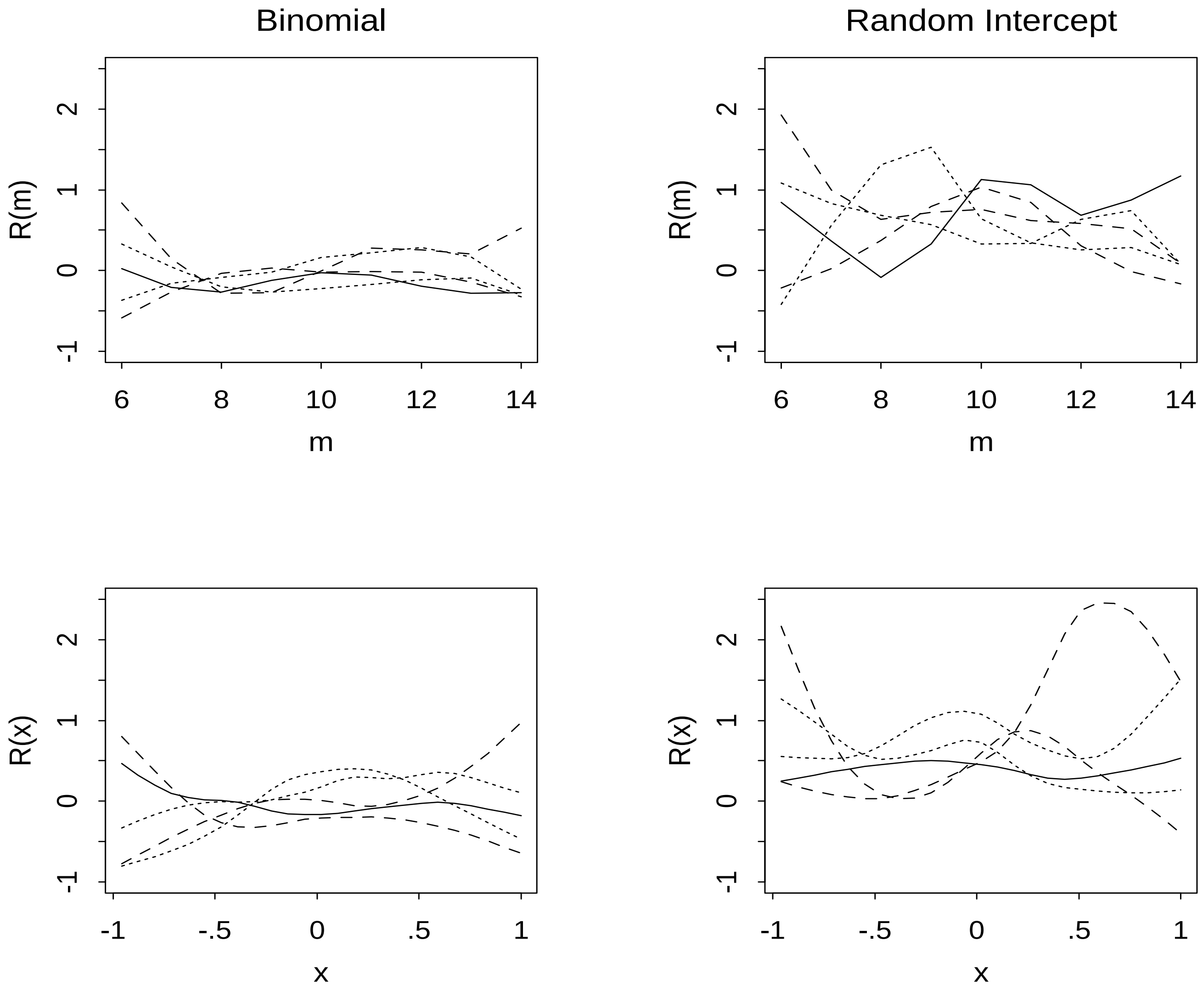



Figure 8:

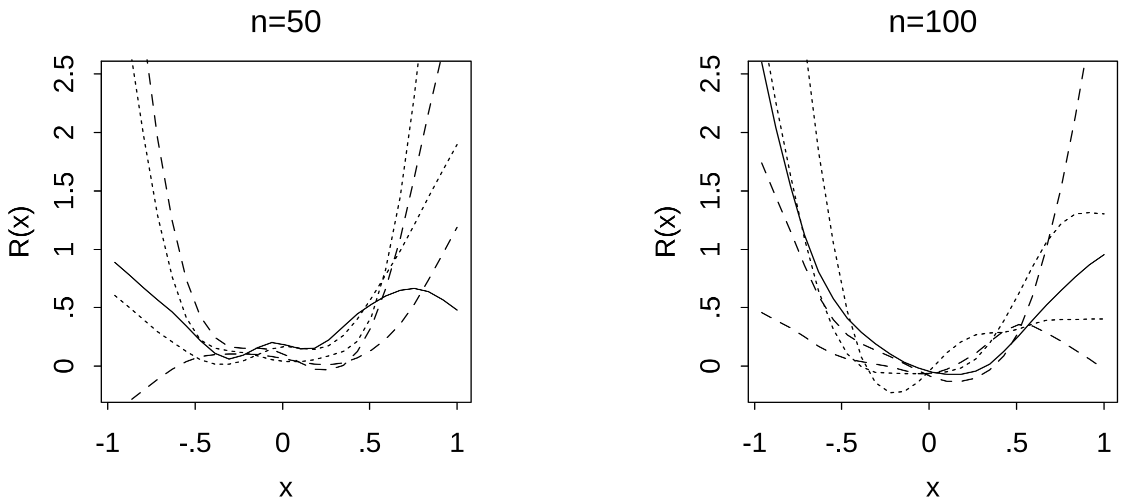



Figure 9: