



*Research article***Covariate-varying dispersion in negative binomial regression: a score-test diagnostic with an application to healthcare utilization****Mohieddine Rahmouni***

Applied College, King Faisal University, Al-Ahsa 31982, Saudi Arabia

* **Correspondence:** Email: mrahmouni@kfu.edu.sa.

Abstract: Negative binomial (NB) regression is the standard model for overdispersed healthcare counts, but it usually assumes a common negative-binomial shape parameter, and hence a common level of extra-Poisson dispersion, for all individuals. This paper develops a score test of this assumption against a log-linear dispersion regression. The test requires only the constant-dispersion fit. The model is identified whether or not the dispersion covariates overlap with the mean covariates, and the expected information is block diagonal between the mean and dispersion parameters. As a result, the score statistic has a closed form and is exactly invariant to the dispersion link. It is asymptotically chi-squared under the null hypothesis, and under local alternatives its noncentrality depends only on a weighted variance of the dispersion covariates. In a Monte Carlo study, the expected-information version keeps its empirical size between 0.03 and 0.06 from $n = 50$ upward, whereas the observed-information and likelihood-ratio versions over-reject in small samples. The method is applied to 14,732 adults in the 2023 Medical Expenditure Panel Survey. Constant dispersion is rejected under all specifications, including design-based tests, and varying dispersion improves cross-validated predictive scores. Complementary diagnostics show that an apparently large insurance effect on dispersion is explained by the effect of insurance on the mean, by excess zeros, and by the choice of variance function. What remains robust is a modest income gradient in dispersion.

Keywords: healthcare utilization; count regression; negative binomial regression; varying dispersion; score test; distributional regression

Mathematics Subject Classification: 62J12, 62F03, 62F05, 62P10

1. Introduction

Healthcare utilization is usually measured by counts: physician and outpatient visits, emergency department attendances, prescription fills, and hospital admissions. These counts are nonnegative and right-skewed, and they typically combine a large mass at or near zero with a long upper tail. Poisson

regression, which forces the conditional variance to equal the conditional mean, is rarely adequate for such data. Negative binomial (NB) regression, which adds a parameter for extra-Poisson variation, is therefore the standard applied model [1, 2].

In its usual form, NB regression uses a single shape parameter for all individuals. Equivalently, the amount of extra-Poisson variation, after conditioning on the mean covariates, is assumed to be the same for everyone. This is an assumption of convenience. Two patients with the same expected number of office visits may differ in how widely their realized use varies around that expectation. Access barriers, interrupted insurance coverage, or unmeasured health shocks are all plausible reasons. Models in which the dispersion has its own regression equation are well established, notably the double generalized linear model (DGLM) [3] and generalized additive models for location, scale, and shape (GAMLSS) [4]. In health-services research, however, they are rarely used, and the constant-dispersion assumption is seldom tested.

A natural diagnostic is a score, or Lagrange multiplier, test of constant dispersion. It needs only the constant-dispersion fit that analysts already compute. Using such a test well, however, requires answers to several questions. Are the mean and dispersion parameters identified when the same covariates enter both equations? How does correlation between mean and dispersion covariates affect precision and power? How accurate is the chi-squared approximation in the sample sizes typical of health studies? And what does a rejection mean when the NB model may also be misspecified in other ways, for example through excess zeros or a variance function that is not quadratic in the mean?

This paper addresses these questions. It shows that the varying-dispersion NB model is identified under full-rank conditions within each equation, with no restriction on how the mean and dispersion covariates are related. The expected information is block diagonal between the mean and dispersion parameters. This gives the score statistic a closed form: a quadratic form in a weighted covariance matrix of the dispersion covariates. The block-diagonal expected information implies that correlation between the two covariate sets does not create asymptotic cross-equation information coupling; its effect on local power enters only through the mean-dependent weights. The statistic is asymptotically chi-squared under the null hypothesis and noncentral chi-squared under local alternatives, and it is exactly invariant to the link used for the dispersion. The analysis also shows that the linear (NB1) and power (NB-P) variance functions are special cases of the quadratic (NB2) model with a mean-dependent shape parameter. This link between dispersion heterogeneity and the choice of variance function is central to interpreting the test in practice.

These results are complemented by an extensive Monte Carlo study and by an application to office-based visits in the 2023 Medical Expenditure Panel Survey (MEPS). The simulations compare the expected-information score test with its observed-information and robust versions and with likelihood-ratio and Wald tests, over a wide range of sample sizes, signal strengths, covariate designs, and misspecified data-generating processes. The application follows a complete workflow: specification of the mean, the score test, zero-modified and alternative variance-function models, survey-design-based inference, effect sizes, and cross-validated prediction. It shows that constant dispersion is clearly rejected, and that the complementary diagnostics are essential for attributing the rejection to its sources.

The paper is organized as follows. Section 2 reviews the related literature. Section 3 introduces the models and establishes identification. Section 4 derives the score test and its asymptotic properties, and Section 5 turns these results into a practical workflow. Section 6 reports the simulation study, Section 7 presents the MEPS application, and Section 8 concludes.

2. Related literature

Regression models in which both the mean and the dispersion depend on covariates were first formalized in the DGLM of Smyth [3], which estimates a mean submodel and a dispersion submodel by joint maximum likelihood. Smyth and Verbyla [5] later proposed adjusted profile likelihood for the dispersion submodel to reduce its small-sample bias. The LMS (lambda-mu-sigma) method of Cole and Green [6], which models location, scale, and a Box-Cox shape parameter, was an early form of distributional modeling. GAMLSS [4, 7, 8] generalized these ideas, allowing every parameter of a parametric response distribution to depend on covariates through smooth additive predictors. It includes several NB parameterizations. Kneib et al. [9] review the wider field of distributional regression.

Within the NB family, the form of the variance function matters as much as the covariates in the dispersion equation. The NB2 model has variance $\mu + \mu^2/\alpha$, quadratic in the mean μ . The NB1 model has variance $\mu(1 + \delta)$, linear in μ [1, 10]. The NB-P model of Greene [11], with variance $\mu + \sigma\mu^p$, nests both. As shown in Section 3.5, NB1 and NB-P can be written as NB2 models whose shape parameter varies with the mean. Testing for covariate-dependent dispersion and choosing the variance function are therefore closely related problems.

Score tests for count models have a long history. Cameron and Trivedi [10] proposed regression-based tests of Poisson overdispersion, Dean [12] derived score statistics for extra-Poisson variation, and Yang et al. [13] developed score tests for overdispersion in zero-inflated models. The score test itself goes back to Rao [14]. It is asymptotically equivalent to the likelihood-ratio (LR) and Wald tests under the null and under contiguous alternatives [15], but it requires only the restricted model, whereas the LR test requires both models and the Wald test the unrestricted one. This property underlies widely used diagnostics such as the Breusch-Pagan test for heteroskedasticity [16] and the regression specification error test (RESET) [17]. Sandwich covariance estimators can also be used for score tests under misspecification when the corresponding score moment conditions remain correctly centered [18, 19]. The test developed here belongs to this family, but it targets the slope of the NB dispersion rather than the adequacy of the NB model relative to the Poisson model.

In health economics, heterogeneity in utilization has been modeled mainly through the mean and the zero process. Deb and Trivedi [20] documented latent-class structure in physician visits. Hurdle and zero-inflated models [21–23] separate the decision to seek care from the amount of care used, and Jones [24] emphasized that the variance specification can change inference on policy-relevant coefficients. Insurance coverage is associated with less unmet need and more stable contact with care [25, 26]. This suggests that insurance and socioeconomic status may be related to the dispersion of utilization as well as to its level. Section 7 examines this hypothesis and shows why it cannot be assessed without also examining the mean, the zero process, and the variance function.

3. Negative binomial models with constant and varying dispersion

This section introduces the NB parameterization used throughout, the varying-dispersion model, and its identification. It closes by relating the model to the NB1 and NB-P variance functions.

3.1. Parameterization

Let Y_i be a count for individual $i = 1, \dots, n$ and let $\tilde{x}_i = (1, x_i^\top)^\top$ contain an intercept and p mean covariates. The NB distribution is used in its mean-shape parameterization:

$$\Pr(Y_i = y \mid \mu_i, \alpha_i) = \frac{\Gamma(y + \alpha_i)}{\Gamma(\alpha_i) y!} \left(\frac{\alpha_i}{\alpha_i + \mu_i} \right)^{\alpha_i} \left(\frac{\mu_i}{\alpha_i + \mu_i} \right)^y, \quad y = 0, 1, 2, \dots, \quad (3.1)$$

with $\mu_i > 0$ and $\alpha_i > 0$. Its moments are

$$\mathbb{E}(Y_i) = \mu_i, \quad \text{Var}(Y_i) = \mu_i + \mu_i^2 / \alpha_i. \quad (3.2)$$

With a constant shape $\alpha_i \equiv \alpha$, Eq (3.1) is the NB2 model. Writing $\phi = 1/\alpha$ expresses the same model in dispersion form, $\text{Var}(Y_i) = \mu_i + \phi \mu_i^2$. In GAMLSS this model is the family NBI (negative binomial type I), with $\sigma = \phi$; despite the Roman numeral, it has the quadratic NB2 variance. The NB1 variance $\mu_i(1 + \delta)$, which is the GAMLSS family NBII, is a different variance function and is discussed in Section 3.5.

Lemma 1 (Poisson-gamma representation). *Let $Y_i \mid v_i \sim \text{Poisson}(\mu_i v_i)$ and $v_i \sim \text{Gamma}(\alpha_i, \alpha_i)$ (shape α_i , rate α_i), so that $\mathbb{E}(v_i) = 1$ and $\text{Var}(v_i) = 1/\alpha_i$. Then the marginal distribution of Y_i is (3.1), with the moments (3.2).*

The proof is given in Appendix A. The lemma clarifies the role of α_i . The quantity $1/\alpha_i$ is the variance of a latent multiplicative frailty v_i , not of the count itself, whose variance is $\mu_i + \mu_i^2/\alpha_i$. For a given mean, a smaller α_i therefore means more extra-Poisson variation. Because the lemma holds for any $\alpha_i > 0$, it also justifies letting α_i depend on covariates.

3.2. The varying-dispersion model

Let z_i be a q -vector of dispersion covariates and let $w_i = (1, z_i^\top)^\top$. The varying-dispersion NB (VD-NB) model is

$$Y_i \mid x_i, z_i \sim \text{NB}(\mu_i, \alpha_i), \quad \log \mu_i = \tilde{x}_i^\top \beta, \quad g(\alpha_i) = w_i^\top \gamma = \eta_0 + z_i^\top \eta, \quad (3.3)$$

where g is a smooth, strictly monotone link (the logarithm unless stated otherwise) and $\theta = (\beta^\top, \gamma^\top)^\top$. The constant-dispersion model corresponds to $\eta = \mathbf{0}_q$, and the hypotheses of interest are

$$H_0 : \eta = \mathbf{0}_q \quad \text{vs.} \quad H_1 : \eta \neq \mathbf{0}_q. \quad (3.4)$$

Under the log link, $\exp(\eta_j)$ multiplies α_i , so a positive η_j means less extra-Poisson variation at a given mean. Such utilization is described as more regular, in the sense of a smaller spread of realized counts around their expectation.

Model (3.3) places no restriction on the relation between x_i and z_i . They may be disjoint, overlapping, or identical, and a variable that enters both equations has separate mean and dispersion coefficients.

3.3. Identification

Proposition 1 (Identification). *Suppose that (3.3) holds with a strictly monotone link g , and that $\mathbb{E}(\tilde{x}_i \tilde{x}_i^\top)$ and $\mathbb{E}(w_i w_i^\top)$ are finite and nonsingular. Then θ is identified: If (β, γ) and (β^*, γ^*) imply the same conditional distribution of Y_i given (x_i, z_i) almost surely, then $\beta = \beta^*$ and $\gamma = \gamma^*$.*

Proof. Equal conditional distributions imply equal conditional means, so $\exp(\tilde{x}_i^\top \beta) = \exp(\tilde{x}_i^\top \beta^*)$ almost surely. Hence $\tilde{x}_i^\top (\beta - \beta^*) = 0$ almost surely, and therefore $(\beta - \beta^*)^\top \mathbb{E}(\tilde{x}_i \tilde{x}_i^\top) (\beta - \beta^*) = 0$. Nonsingularity gives $\beta = \beta^*$, so μ_i is the same under both parameter values. Equal conditional variances then give $\mu_i^2 / \alpha_i = \mu_i^2 / \alpha_i^*$. Since $\mu_i > 0$, this means $\alpha_i = \alpha_i^*$, and hence $w_i^\top \gamma = w_i^\top \gamma^*$ almost surely. The same argument with $\mathbb{E}(w_i w_i^\top)$ gives $\gamma = \gamma^*$. $\square$

Identification thus requires full rank within each design, but imposes no condition on the relation between the mean and dispersion covariates. In particular, the specification $z_i = x_i$ is identified: The mean determines β , and the variance around that mean determines γ .

Remark 1 (Exclusion restrictions are testable). *A common modeling choice places some variables only in the dispersion equation, assuming that they do not shift the mean. By Proposition 1, this exclusion restriction can be tested within the NB framework: The variables are added to the mean equation, and the joint significance of their coefficients is assessed by an LR, Wald, or score test. Section 7.2 applies this test to the MEPS data.*

3.4. Choice of dispersion covariates

The vector z_i represents a hypothesis about the sources of dispersion, and three types of specification are useful in practice (Table 1). In the shared specification ($z_i = x_i$), the test asks whether any mean predictor is also associated with dispersion, which makes it an omnibus check. In the subset specification ($z_i \subseteq x_i$), prior knowledge restricts dispersion to a few mean predictors. In the distinct specification ($z_i \not\subseteq x_i$), dispersion is hypothesized to depend on variables outside the mean equation; whether those variables should also enter the mean can be tested as in Remark 1. Specifying z_i in advance keeps the chi-squared reference distribution valid. A search over many candidate vectors would require a correction for multiple testing.

Table 1. Dispersion covariate specifications (df : degrees of freedom of the score test).

Specification	Relationship	Score-test df	Primary use
Shared	$z_i = x_i$	$q = p$	Omnibus check; also sensitive to the variance function (Section 3.5)
Subset	$z_i \subseteq x_i$	$q < p$	Dispersion restricted to some mean predictors
Distinct	$z_i \not\subseteq x_i$	$q = \dim(z_i)$	Dispersion hypothesized to depend on variables outside the mean equation

3.5. NB1 and NB-P as special cases of a varying NB2 shape

Lemma 1 also shows how other members of the NB family arise. Setting $\alpha_i = \mu_i / \delta$ in (3.2) gives $\text{Var}(Y_i) = \mu_i(1 + \delta)$, the NB1 model. More generally, $\alpha_i = \mu_i^{2-P} / \sigma$ gives the NB-P variance $\mu_i + \sigma \mu_i^P$ [11]. On the log scale,

$$\log \alpha_i = -\log \sigma + (2 - P) \tilde{x}_i^\top \beta, \quad (3.5)$$

which is a VD-NB2 model with $z_i = x_i$ and $\eta = (2 - P)\beta$ (excluding the intercept). This has a direct practical consequence. If the true variance function is closer to NB1 than to NB2, the constant-dispersion NB2 model is misspecified in a direction correlated with $\log \mu_i$. A score test with dispersion covariates related to the mean will then tend to reject, more often as n grows, even without any substantive dispersion heterogeneity. The shared specification is especially sensitive to this. Comparing NB2 fits with NB1 or NB-P fits is therefore an integral part of interpreting the test (Section 5).

4. Score test for varying dispersion

This section derives the score and information of model (3.3), constructs the score statistic, and establishes its asymptotic distribution. It also examines how the relation between x_i and z_i affects the test.

4.1. Score and Hessian

Let $s_i = \alpha_i + \mu_i$, $a_i = \log \mu_i$, and $b_i = \log \alpha_i$. The log-likelihood contribution is

$$\ell_i(\theta) = \log \Gamma(Y_i + \alpha_i) - \log \Gamma(\alpha_i) - \log(Y_i!) + \alpha_i \log \frac{\alpha_i}{s_i} + Y_i \log \frac{\mu_i}{s_i}. \quad (4.1)$$

With ψ and ψ' denoting the digamma and trigamma functions, the first derivatives with respect to the linear predictors are

$$\frac{\partial \ell_i}{\partial a_i} = \frac{\alpha_i(Y_i - \mu_i)}{s_i}, \quad (4.2)$$

$$\frac{\partial \ell_i}{\partial b_i} = \alpha_i \ell_{\alpha,i}, \quad (4.3)$$

$$\ell_{\alpha,i} = \psi(Y_i + \alpha_i) - \psi(\alpha_i) + \log\left(\frac{\alpha_i}{s_i}\right) + 1 - \frac{Y_i + \alpha_i}{s_i}.$$

The score vectors are $U_\beta = \sum_i \tilde{x}_i \partial \ell_i / \partial a_i$ and $U_\gamma = \sum_i w_i \partial \ell_i / \partial b_i$. The second derivatives are

$$\frac{\partial^2 \ell_i}{\partial a_i^2} = -\frac{\alpha_i \mu_i (\alpha_i + Y_i)}{s_i^2}, \quad \frac{\partial^2 \ell_i}{\partial a_i \partial b_i} = \frac{\alpha_i \mu_i (Y_i - \mu_i)}{s_i^2}, \quad (4.4)$$

$$\frac{\partial^2 \ell_i}{\partial b_i^2} = \alpha_i \ell_{\alpha,i} + \alpha_i^2 \ell_{\alpha\alpha,i}, \quad \ell_{\alpha\alpha,i} = \psi'(Y_i + \alpha_i) - \psi'(\alpha_i) + \frac{1}{\alpha_i} - \frac{1}{s_i} - \frac{\mu_i - Y_i}{s_i^2}. \quad (4.5)$$

The observed Hessian with respect to θ follows by the chain rule, for example,

$$\frac{\partial^2 \ell}{\partial \beta \partial \gamma^\top} = \sum_i \tilde{x}_i w_i^\top \frac{\partial^2 \ell_i}{\partial a_i \partial b_i}.$$

Appendix B treats general links.

4.2. Expected information

Lemma 2 (Block-diagonal information). *Let $\pi_k(\mu, \alpha) = \Pr(Y > k)$ for $Y \sim NB(\mu, \alpha)$ and define*

$$i_{\alpha\alpha}(\mu, \alpha) = \sum_{k=0}^{\infty} \frac{\pi_k(\mu, \alpha)}{(\alpha + k)^2} - \frac{1}{\alpha} + \frac{1}{\alpha + \mu}, \quad c(\mu, \alpha) = \alpha^2 i_{\alpha\alpha}(\mu, \alpha). \quad (4.6)$$

Under the log link, the Fisher information for $\theta = (\beta, \gamma)$ from observation i is

$$\mathcal{I}_i(\theta) = \begin{pmatrix} \frac{\alpha_i \mu_i}{\alpha_i + \mu_i} \tilde{x}_i \tilde{x}_i^\top & \mathbf{0} \\ \mathbf{0} & c(\mu_i, \alpha_i) w_i w_i^\top \end{pmatrix}. \quad (4.7)$$

The weight $c(\mu, \alpha)$ is positive for $\mu > 0$, tends to 0 as $\mu \rightarrow 0$, and tends to $\alpha^2\{\psi'(\alpha) - 1/\alpha\}$ as $\mu \rightarrow \infty$.

The proof is given in Appendix B. It rests on two facts: $\mathbb{E}(Y_i - \mu_i) = 0$ removes the cross term in (4.4), and $\mathbb{E}\{\psi'(Y + \alpha)\} - \psi'(\alpha) = -\sum_k \pi_k/(\alpha + k)^2$. The series converges geometrically and can be evaluated to machine precision. Numerically, $c(\mu, \alpha)$ is increasing in μ over the range examined ($0.2 \leq \alpha \leq 5$ and $0.01 \leq \mu \leq 200$). Observations with very small expected counts therefore carry little information about dispersion, since a count that is almost always zero reveals little about its spread.

4.3. The test statistic

Let $\hat{\theta}_0 = (\hat{\beta}^\top, \hat{\eta}_0, \mathbf{0}^\top)^\top$ denote the constant-dispersion maximum likelihood estimator (MLE), with $\hat{\alpha} = e^{\hat{\eta}_0}$ and $\hat{\mu}_i = \exp(\tilde{x}_i^\top \hat{\beta})$. Define $\hat{u}_i = \hat{\alpha} \ell_{\alpha,i}(\hat{\theta}_0)$ and $\hat{c}_i = c(\hat{\mu}_i, \hat{\alpha})$. The score for η is $U_\eta = \sum_i z_i \hat{u}_i$. Because the first-order condition for η_0 gives $\sum_i \hat{u}_i = 0$, it can be written $U_\eta = \sum_i (z_i - \bar{z}_c) \hat{u}_i$ for any centering vector $\bar{z}_c$. With nuisance parameters $\lambda = (\beta^\top, \eta_0)^\top$, the efficient information for η is the Schur complement $\mathcal{I}_{\eta\eta\cdot\lambda} = \mathcal{I}_{\eta\eta} - \mathcal{I}_{\eta\lambda} \mathcal{I}_{\lambda\lambda}^{-1} \mathcal{I}_{\lambda\eta}$. By Lemma 2, the β block drops out, and

$$\hat{\mathcal{I}}_{\eta\eta\cdot\lambda} = \sum_{i=1}^n \hat{c}_i (z_i - \bar{z}_c)(z_i - \bar{z}_c)^\top, \quad \bar{z}_c = \frac{\sum_i \hat{c}_i z_i}{\sum_i \hat{c}_i}. \quad (4.8)$$

The score statistic is

$$S = U_\eta^\top \hat{\mathcal{I}}_{\eta\eta\cdot\lambda}^{-1} U_\eta, \quad (4.9)$$

a quadratic form in the $\hat{c}$ -weighted covariance matrix of the dispersion covariates. It is referred to as the expected-information score statistic and is the version recommended here. Three alternatives are also considered:

- (i) the observed-information version, which replaces (4.8) by the Schur complement of the observed Hessian (4.4) and (4.5);
- (ii) the robust version, which replaces $\mathcal{I}_{\eta\eta\cdot\lambda}$ by $\sum_i \tilde{U}_i \tilde{U}_i^\top$, where $\tilde{U}_i = U_{\eta,i} - \mathcal{I}_{\eta\lambda} \mathcal{I}_{\lambda\lambda}^{-1} U_{\lambda,i}$ are the efficient score contributions [19]. This adjustment estimates the variance of the efficient score without imposing the information identity. It can improve calibration under some forms of misspecification, but it is not a general safeguard against an incorrectly centered dispersion score;
- (iii) a parametric bootstrap calibration, which draws B samples from the fitted null model, recomputes S on each, and reports $\hat{p} = \{1 + \#(S_b^* \geq S)\}/(B + 1)$ [27, 28].

4.4. Asymptotic distribution

The asymptotic results require the following conditions.

Assumption 1.

(A1) (Y_i, x_i, z_i) , $i = 1, \dots, n$, are independent and identically distributed, and $Y_i \mid (x_i, z_i)$ follows (3.3) with the log link.

(A2) The covariates have compact support.

(A3) The true value θ_0 lies in the interior of a compact parameter space Θ .

(A4) $\mathbb{E}(\tilde{x}_i \tilde{x}_i^\top)$ and $\mathbb{E}(w_i w_i^\top)$ are nonsingular.

Condition (A4) excludes collinearity within the mean design and within the dispersion design, but places no restriction on the relation between x_i and z_i . Under (A2) and (A3), μ_i and α_i are bounded away from 0 and ∞ uniformly on Θ . Under (A4), and because $c(\mu, \alpha) > 0$ for $\mu > 0$, the efficient information $\mathcal{I}^* = \mathbb{E}\{c_i(z_i - m_c)(z_i - m_c)^\top\}$, with $m_c = \mathbb{E}(c_i z_i)/\mathbb{E}(c_i)$, is positive definite.

Theorem 1. Under Assumption 1:

- (a) under H_0 , $S \xrightarrow{d} \chi_q^2$, and the same limit holds for the observed-information and robust versions;
- (b) under the local alternatives $\eta_n = h/\sqrt{n}$, $S \xrightarrow{d} \chi_q^2(\kappa)$ with noncentrality

$$\kappa = h^\top \mathcal{I}^* h = \mathbb{E}[c(\mu_i, \alpha) \{(z_i - m_c)^\top h\}^2]; \quad (4.10)$$

- (c) the statistic S in (4.9) is exactly invariant to the choice of a smooth, strictly monotone link g in (3.3).

The proof is given in Appendix C. Part (c) holds because, at $\eta = \mathbf{0}$, the derivative $\partial \alpha_i / \partial \eta = z_i / g'(\alpha)$ is the same scalar multiple of z_i for every observation. The score and the expected information are therefore rescaled by the same factor, which cancels in S . The observed-information version is only asymptotically invariant, because the observed Hessian contains the term $\ell_{\alpha,i} \partial^2 \alpha_i / \partial \eta \partial \eta^\top$, which depends on g'' . Section 7.8 shows that the difference can be large in practice.

4.5. Correlation between mean and dispersion covariates

Lemma 2 and Theorem 1 also show how correlation between x_i and z_i affects estimation and testing. Because the expected information is block diagonal, $\hat{\beta}$ and $\hat{\gamma}$ are asymptotically independent. Correlation between the two sets of covariates does not inflate the standard errors of one equation through the other, as collinearity within a single linear predictor would. The observed cross block,

$$\sum_i \tilde{x}_i w_i^\top \frac{\alpha_i \mu_i (Y_i - \mu_i)}{s_i^2}$$

has mean zero and is $O_p(\sqrt{n})$.

For the same reason, estimating β costs no information about η asymptotically. The only nuisance adjustment in (4.8) is centering z_i at its c -weighted mean, which accounts for the estimation of η_0 . By (4.10), power is determined by the c -weighted variance of the dispersion index $z_i^\top h$. Correlation with x_i enters only through the weights $c(\mu_i, \alpha)$, which vary with $\mu_i = \exp(\tilde{x}_i^\top \beta)$. Two designs with the same weighted variance of $z_i^\top \eta$ therefore have essentially the same power, whether or not z_i is correlated with x_i . Section 6 confirms this numerically.

4.6. Relation to the likelihood-ratio and Wald tests

The LR statistic $2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_0)\}$ and the Wald statistic $\hat{\eta}^\top \widehat{\text{Var}}(\hat{\eta})^{-1} \hat{\eta}$ have the same limits as S under H_0 and under local alternatives [15]. Both require fitting the varying-dispersion model. They differ from S when the log-likelihood is far from quadratic in η , as in small samples or for large departures from the null. Their finite-sample behavior is compared with that of S in Section 6.

5. Practical workflow

The results above suggest a sequence of steps for applied work (Figure 1). The score test is the central screening step. Because a rejection can reflect omitted mean effects, a separate zero-generating process, or a non-quadratic variance function (Section 3.5), the workflow also includes checks of these alternatives and a report of the magnitude of any dispersion effects.

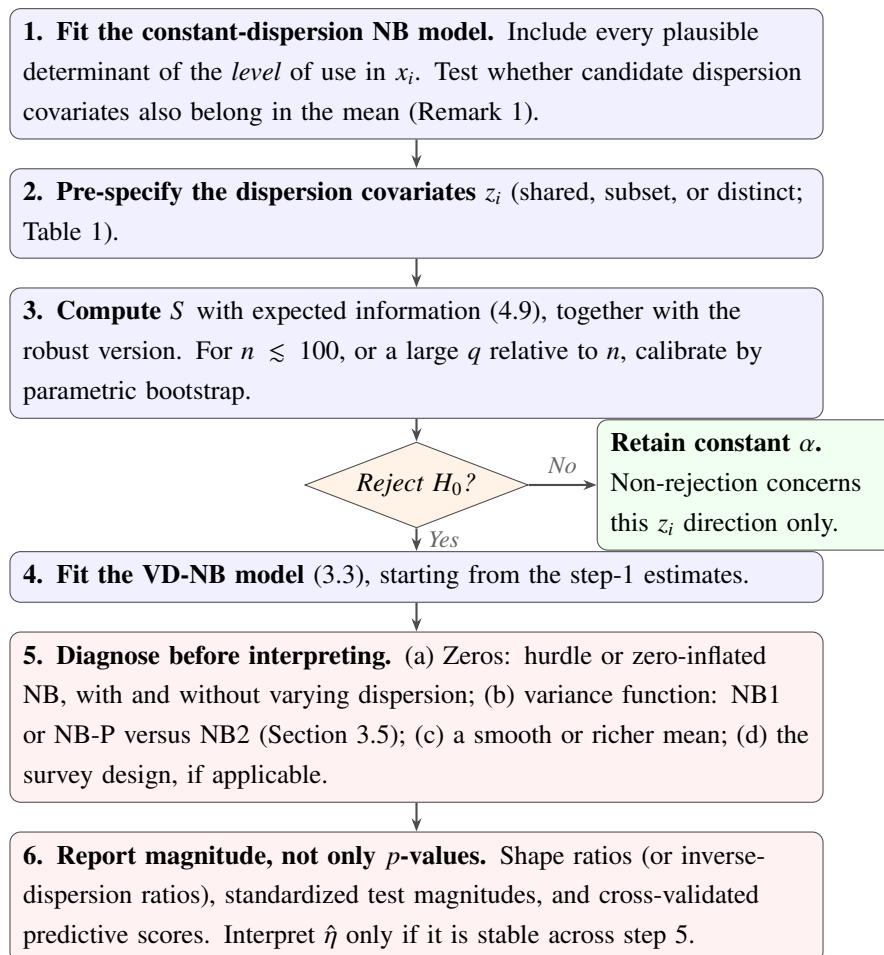


Figure 1. Workflow for assessing and interpreting covariate-varying dispersion in NB regression.

The simulation study in Section 6 gives guidance on sample size for designs with two dispersion covariates. The expected-information score test keeps its size within the liberal band $[0.025, 0.075]$ of

Bradley [29] for every $n \geq 50$ examined. The LR test needs about $n \geq 100$ for the same accuracy, and the observed-information score test about $n \geq 200$. In most settings, power rather than size is the binding constraint. About 80% power requires $n \approx 500$ when the dispersion index $z_i^\top \eta$ has standard deviation 0.5, and $n \approx 1,000$ when its standard deviation is 0.36. Weak heterogeneity, with a standard deviation of 0.18, is not reliably detected even at $n = 2,000$. Because the information about α vanishes as $\mu \rightarrow 0$ (Lemma 2), samples dominated by very small expected counts need larger n . In such cases the parametric bootstrap is a convenient safeguard, since it requires only null-model fits.

Continuous covariates are best centered and scaled, particularly in the dispersion equation. This improves the conditioning of the information matrix and makes the dispersion slopes comparable. The expected-information statistic itself is invariant to affine transformations of z_i .

6. Simulation study

The simulation study examines the finite-sample size and power of the score test and its alternatives, the role of the dispersion design and of correlation between covariates, the value of bootstrap calibration, and the behavior of the tests under misspecified data-generating processes.

6.1. Design

Data were generated from (3.3) with $\beta = (0.5, 0.4, -0.3)^\top$, $\eta_0 = 0.5$, and $x_{1i}, x_{2i} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$. The median of μ_i is $e^{0.5} \approx 1.65$. Four signal levels,

$$(\eta_1, \eta_2) \in \{(0, 0), (0.15, 0.10), (0.40, 0.30), (0.60, 0.45)\},$$

labeled null, weak, moderate, and strong, were combined with seven sample sizes,

$$n \in \{50, 100, 200, 300, 500, 1,000, 2,000\}.$$

Three dispersion designs were used. In the shared design, $z_i = x_i$. In the distinct design, $z_{1i} \sim \text{Bernoulli}(0.5)$ and $z_{2i} \sim \mathcal{N}(0, 1)$, independently of x_i . The rescaled distinct design is the same, except that z_{1i} is standardized to unit variance, so that $\text{Var}(z_i^\top \eta)$ equals its value in the shared design. A fourth design isolates the effect of correlation between the mean and dispersion covariates by setting

$$z_{ji} = \rho x_{ji} + \sqrt{1 - \rho^2} e_{ji},$$

with $e_{ji} \sim \mathcal{N}(0, 1)$ and $\rho \in \{-1, -0.75, \dots, 1\}$. In every replication, the expected-information, observed-information, and robust score tests were computed, together with the LR and Wald tests. Each cell of the main design has $R = 1,000$ replications (84,000 in total), so the Monte Carlo standard error (MCSE) of a rejection rate near 0.05 is about 0.007.

Table 2 shows how much the conditional variance varies in each design. At the moderate signal, the 5th to 95th percentile range of α_i is 0.72–3.76 in the shared design and 1.11–3.64 in the distinct design. The standard deviation of the dispersion index $z_i^\top \eta$ is 0.50 and 0.36, respectively. The difference arises because the binary z_{1i} has variance 0.25, which compresses its dispersion effect. The rescaled design removes this difference and allows the two designs to be compared at equal signal strength.

Table 2. Dispersion heterogeneity in the simulation designs (computed from 4×10^5 draws). Var_0 is the variance under constant $\alpha = e^{0.5}$ at the same μ_i . The last column gives the 5th–95th percentile range of $\text{Var}(Y | \mu, \alpha_i)$ at the median mean $\tilde{\mu} = 1.65$; under the null, $\alpha_i \equiv 1.65$ and this variance equals 3.30.

Design	Signal	SD ($z_i^\top \eta$)	α_i percentiles			$\text{Var}(Y_i x_i, z_i) / \text{Var}_0$			Var at $\tilde{\mu}$: 5th–95th
			5th	50th	95th	5th	50th	95th	
Shared	Weak	0.18	1.23	1.65	2.22	0.85	1.00	1.15	2.88–3.87
Shared	Moderate	0.50	0.72	1.65	3.76	0.68	1.00	1.58	2.37–5.41
Shared	Strong	0.75	0.48	1.65	5.68	0.59	1.00	2.10	2.13–7.32
Distinct	Weak	0.12	1.45	1.78	2.18	0.87	0.97	1.07	2.89–3.52
Distinct	Moderate	0.36	1.11	2.01	3.64	0.71	0.92	1.24	2.39–4.08
Distinct	Strong	0.54	0.92	2.23	5.41	0.62	0.88	1.40	2.15–4.62
Distinct, rescaled	Weak	0.18	1.25	1.65	2.18	0.87	1.00	1.17	2.89–3.82
Distinct, rescaled	Moderate	0.50	0.75	1.65	3.61	0.71	1.00	1.61	2.40–5.26
Distinct, rescaled	Strong	0.75	0.51	1.65	5.37	0.63	1.00	2.15	2.16–7.03

6.2. Size

Figure 2 and Table A1 in Appendix D report the empirical size of the five tests. The expected-information score test has rejection rates between 0.029 and 0.059 in all 21 null cells. It is mildly conservative at $n = 50$ and otherwise close to 0.05. The observed-information version over-rejects in small samples, with rates of 0.091–0.115 at $n = 50$ and 0.071–0.097 at $n = 100$. The LR test also over-rejects at $n = 50$ (0.098–0.112). The Wald test is conservative up to $n \approx 300$, and the robust score test is conservative for $n \leq 100$. From $n = 500$ onward, all five tests are close to the nominal level. The size distortion of the observed-information version is due to the observed Hessian itself, not to how it is computed: Analytic and central-difference Hessians give statistics that agree to within 0.1% in the MEPS data.

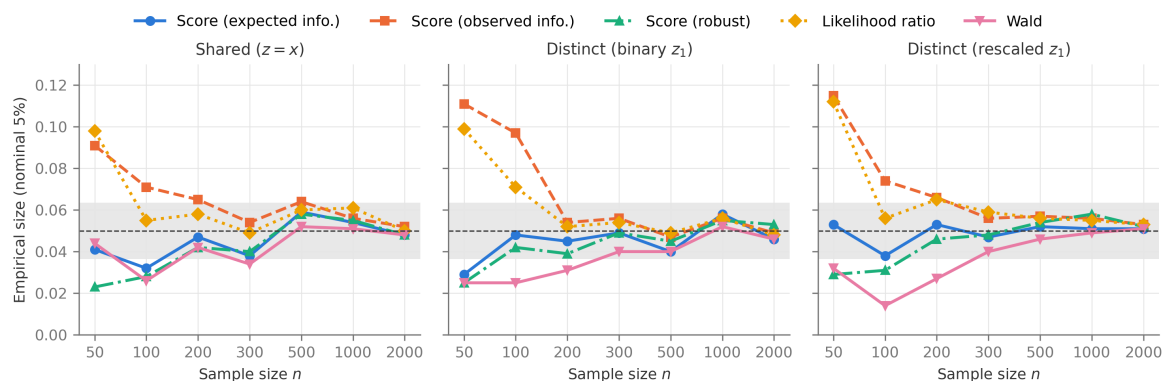


Figure 2. Empirical size at the 5% level by test, design, and sample size ($R = 1,000$ per point). The shaded band is 0.05 ± 1.96 MCSE.

6.3. Power

Figure 3 shows that the power of the expected-information score test increases with n and with signal strength. When $\text{Var}(z_i^\top \eta)$ is equal, the shared and rescaled distinct designs have practically identical power, 0.873 and 0.880 at $n = 500$ for the moderate signal. The distinct design with a binary covariate reaches only 0.581, because its signal is smaller. This agrees with Section 4.5: Power is driven by the variance of the dispersion index, not by whether z_i overlaps with x_i . The LR test rejects slightly more often than the expected-information score test at a given n (Table A2), but part of this apparent advantage reflects its over-rejection under the null in small samples.

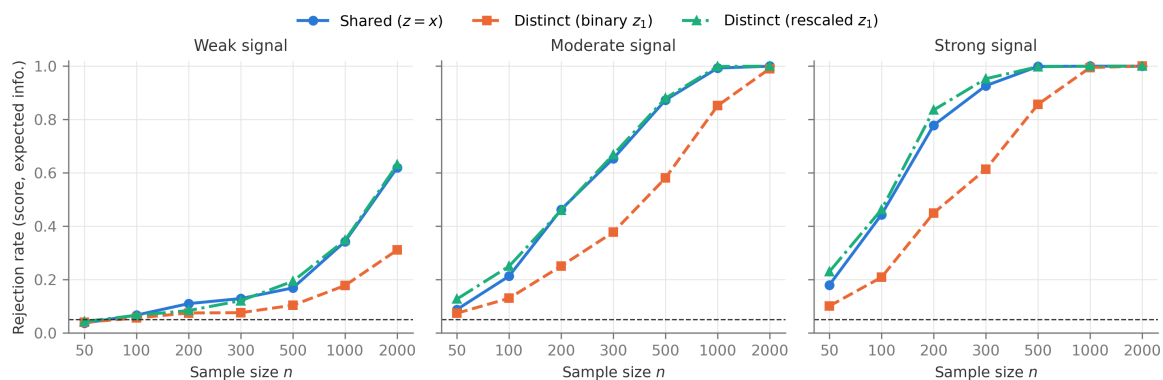


Figure 3. Power of the expected-information score test by signal level, design, and sample size ($R = 1,000$ per point). The dashed line marks 0.05.

Figure 4 gives a direct check of the role of correlation. Across $\rho \in [-1, 1]$ at $n = 300$, the size of the expected-information score test lies between 0.041 and 0.060. Its power at the moderate signal lies between 0.645 and 0.697 at $n = 300$ and between 0.991 and 0.998 at $n = 1,000$, with no trend in ρ .

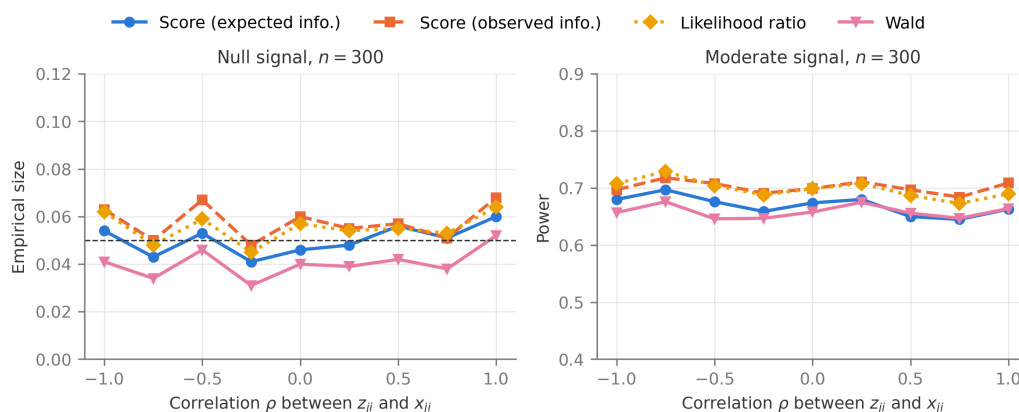


Figure 4. Size (left) and power at the moderate signal (right) as functions of the correlation ρ between z_{ji} and x_{ji} , with $n = 300$ and $R = 1,000$ per point (MCSE about 0.007 for size and 0.015 for power).

6.4. Bootstrap calibration

For small samples, the parametric bootstrap offers an alternative to the chi-squared reference distribution. Table 3 compares asymptotic and bootstrap ($B = 199$) p -values for $n \leq 200$, with $R = 500$ replications per cell. Under the null, the bootstrap rejection rates lie between 0.038 and 0.062, compared with 0.030–0.054 for the asymptotic expected-information test and 0.136 at worst for the observed-information test. Under the moderate alternative, the bootstrap and asymptotic expected-information tests have similar power (differences between -0.030 and $+0.036$). At $n = 50$ the bootstrap removes most of the mild conservativeness of the asymptotic test (0.038 versus 0.030 in the shared design and 0.050 versus 0.044 in the distinct design). The bootstrap is therefore a safe choice in small samples, but with the expected information it is rarely necessary.

Table 3. Rejection rates of asymptotic and parametric-bootstrap ($B = 199$) score tests at the 5% level, with $R = 500$ replications per cell (MCSE about 0.010 under the null).

Signal	Design	n	Score, expected	Score, observed	Score, bootstrap
Null	Shared	50	0.030	0.094	0.038
		100	0.038	0.066	0.040
		200	0.042	0.074	0.050
	Distinct	50	0.044	0.136	0.050
		100	0.054	0.100	0.054
		200	0.050	0.066	0.062
Moderate	Shared	50	0.096	0.210	0.132
		100	0.214	0.304	0.218
		200	0.462	0.526	0.432
	Distinct	50	0.078	0.172	0.096
		100	0.142	0.188	0.140
		200	0.250	0.290	0.240

6.5. Misspecified data-generating processes

A rejection of H_0 is informative only if the test does not also respond strongly to other departures from the constant-dispersion NB2 model. To assess this, data were generated with no dispersion covariates from two alternative processes. The first is a zero-inflated NB model with 15% structural zeros, with the NB mean rescaled so that $\mathbb{E}(Y_i)$ is unchanged. The second is a Poisson-lognormal mixture with the same frailty variance $1/\alpha$ but heavier tails. Table 4 reports the results.

Table 4. Rejection rates at the 5% level under misspecified data-generating processes with no dispersion covariates ($R = 1,000$ per cell). All five tests are of $H_0 : \eta = \mathbf{0}_q$: score tests with expected, observed, and robust (sandwich) information, the likelihood-ratio (LR) test, and the Wald test.

DGP	Design	n	Score (expected)	Score (observed)	Score (robust)	LR	Wald
Zero-inflated NB ($\pi = 0.15$)	Distinct	300	0.046	0.059	0.050	0.060	0.051
		1,000	0.043	0.055	0.059	0.055	0.052
		2,000	0.032	0.040	0.038	0.040	0.038
	Shared	300	0.055	0.058	0.051	0.064	0.046
		1,000	0.065	0.066	0.062	0.073	0.059
		2,000	0.088	0.093	0.091	0.095	0.093
Poisson-lognormal	Distinct	300	0.080	0.084	0.055	0.085	0.059
		1,000	0.092	0.084	0.050	0.086	0.077
		2,000	0.081	0.072	0.045	0.073	0.069
	Shared	300	0.084	0.105	0.053	0.091	0.079
		1,000	0.091	0.095	0.056	0.086	0.083
		2,000	0.118	0.123	0.082	0.113	0.122

When the dispersion covariates are independent of x_i (distinct design), constant zero-inflation does not produce spurious rejections. When $z_i = x_i$, rejection rates increase with n , reaching 0.088–0.095 at $n = 2,000$. In this case the test detects a misspecification that is correlated with μ_i , as anticipated in Section 3.5. Under the Poisson-lognormal process, all model-based tests over-reject (0.059–0.123). The robust score test stays at 0.045–0.056, except for the shared design at $n = 2,000$ (0.082). A rejection by the model-based tests therefore shows that the constant-dispersion NB2 model is inadequate in the direction of z_i . Attributing it to dispersion heterogeneity requires the additional checks in step 5 of Figure 1. The MEPS application illustrates these checks.

7. Application: office-based visits in MEPS 2023

This section applies the workflow of Figure 1 to annual office-based visits of U.S. adults. The analysis asks whether demographic, socioeconomic, and insurance variables are associated with the dispersion of utilization, and whether any such association survives the diagnostic checks.

7.1. Data

The data come from the 2023 MEPS Full-Year Consolidated public-use file (HC-251) released by the Agency for Healthcare Research and Quality (AHRQ) [30]. MEPS is a nationally representative survey of the U.S. civilian non-institutionalized population, with a stratified, clustered design. The analysis sample contains 14,732 adults aged 18 or older with complete information on utilization, health status, insurance, and socioeconomic variables. The outcome is the annual number of office-based provider visits (OBTOTV23). It has mean 7.97, median 3, standard deviation 14.28, and

maximum 419, and 24.5% of adults report no visits. Zeros are much more common among the uninsured (72.9%) than among adults with private insurance (20.9%) or public coverage only (19.9%), who make up 7.5%, 60.6%, and 31.9% of the sample, respectively. Table 5 summarizes the sample. The design-based analysis in Section 7.9 uses the person weights (PERWT23F), variance strata (VARSTR), and primary sampling units (PSUs; VARPSU). The sample spans 105 strata and 262 PSUs, giving 157 design degrees of freedom. The 281 records with a zero person weight are retained for variance estimation only.

Table 5. MEPS 2023 analysis sample.

Quantity	Value	Quantity	Value
Rows in HC-251	18,919	Mean age (years)	51.9
Adults aged 18+	15,123	Share female	0.534
Complete-case sample (n)	14,732	Any private insurance	0.606
Mean office-based visits	7.967	Public-only insurance	0.319
Median	3	Uninsured	0.075
Standard deviation	14.284	Hypertension	0.373
Share with zero visits	0.245	Diabetes	0.139
Maximum	419	Asthma	0.139
Strata / PSUs	105 / 262	Coronary heart disease	0.057

7.2. Specification

The clinical covariates x_i are age in decades, self-rated physical and mental health (0–4, higher is worse), and indicators for hypertension, coronary heart disease, diabetes, and asthma [20, 24]. The candidate dispersion covariates z_i are female sex, years of education, four income-category indicators (poor is the reference category), and indicators for public-only and any private insurance (uninsured is the reference category).

A tempting specification places z_i only in the dispersion equation, on the view that clinical need determines the expected number of visits while access and resources determine its variability. This exclusion restriction is not plausible a priori. Insurance lowers the price of a visit, and education and income affect both health needs and access to care. Following Remark 1, the restriction was tested by adding z_i to the mean of the constant-dispersion model. The log-likelihood increases by 882.0, and the LR statistic of 1,764.0 on 8 degrees of freedom rejects the restriction decisively. The insurance rate ratios in the mean are 3.97 (public only) and 3.71 (any private), relative to the uninsured. The main specification is therefore

$$\log \mu_i = \beta_0 + x_i^\top \beta_x + z_i^\top \beta_z, \quad \log \alpha_i = \eta_0 + z_i^\top \eta. \quad (7.1)$$

It asks whether the socioeconomic and insurance variables are associated with dispersion once their association with the level of use has been accounted for. For comparison, results are also reported for the restricted-mean specification, with mean x_i and dispersion z_i . Intermediate specifications, which move insurance and income into the mean one block at a time, are examined in Section 7.5.

7.3. Score tests and model comparison

Table 6 reports the tests of $H_0 : \eta = \mathbf{0}$. Under the restricted-mean specification, the expected-information score statistic is 1,109.5 and the observed-information statistic is 1,218.0. Under the main specification (7.1), all statistics are considerably smaller: The expected-information score statistic is 660.3, the LR statistic 573.8, and the robust score statistic 164.4. All remain far beyond the χ^2_8 critical value. Table 7 compares the models using the Akaike and Bayesian information criteria (AIC and BIC). Both criteria prefer the main specification to the restricted-mean one, and both prefer varying to constant dispersion within each mean specification.

Table 6. Tests of constant dispersion in the MEPS data. The standardized test magnitude $s = \sqrt{\text{statistic}/n}$ is reported to put statistics on a common per-observation scale. It is not interpreted using Cohen's chi-square benchmarks. All p -values are below 10^{-20} .

Test of $H_0 : \eta = \mathbf{0}$ (z_i : 8 variables)	Restricted mean (x only)		Main specification (mean $x + z$)	
	Statistic	s	Statistic	s
Score, expected information (recommended)	1,109.5	0.274	660.3	0.212
Score, observed information	1,218.0	0.288	602.9	0.202
Score, robust (sandwich)	789.1	0.231	164.4	0.106
Likelihood ratio	1,259.1	0.292	573.8	0.197
Wald	1,008.2	0.262	553.5	0.194
Design-based score (weights, strata, PSUs)	250.4	0.130	113.5	0.088
Design-adjusted Wald F (8, 150)	49.2	—	30.6	—
Score, $z_i = \text{all mean covariates}$ (expected info.)	1,067.7 ($df = 7$)	0.269	1,477.1 ($df = 15$)	0.317

Table 7. Log-likelihood and information criteria.

Model	Parameters	Log-likelihood	AIC	BIC
NB, mean x , constant α	9	−43,497.8	87,013.6	87,081.9
VD-NB, mean x , $z_i = x_i$ (shared)	16	−42,954.6	85,941.1	86,062.7
VD-NB, mean x , dispersion z (restricted mean)	17	−42,868.2	85,770.4	85,899.6
NB, mean $x + z$, constant α	17	−42,615.8	85,265.5	85,394.7
VD-NB, mean $x + z$, dispersion z (main)	25	−42,328.8	84,707.7	84,897.6
VD-NB, mean $x + z$, dispersion $x + z$	32	−41,903.6	83,871.3	84,114.4

With $n = 14,732$, even small departures from H_0 produce very small p -values [31], so measures of magnitude are needed. The standardized test magnitude $s = \sqrt{\text{statistic}/n}$ decreases from 0.27 under the restricted-mean specification to 0.21 under the main specification, and to 0.11 for the robust statistic. This quantity is used only as a scale-free summary of test magnitude; the more interpretable effect measures are the fitted-shape percentiles, implied variance ratios, and predictive score differences reported below. Under the main specification, varying dispersion raises the log-likelihood by 0.019 per

observation. The 10th and 90th percentiles of $\hat{\alpha}_i$ are 0.40 and 0.78. At the median fitted mean of 7.06, this corresponds to a variance ratio of 1.85 between these individuals, or 1.36 in standard deviations. Figure 5(a) shows that, for most of the sample, the implied standard-deviation curves are close to the constant-dispersion curve. The heterogeneity is concentrated among the uninsured, whose median $\hat{\alpha}_i$ is 0.14, compared with 0.61–0.63 for the insured. Dispersion thus differs strongly for a small group and only modestly for the rest of the population.

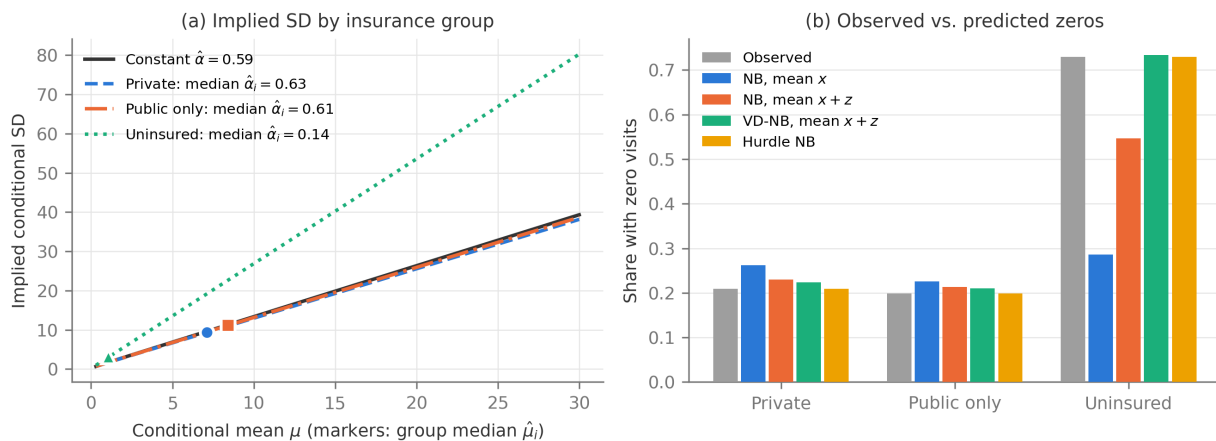


Figure 5. (a) Implied conditional standard deviation under constant dispersion and at the median $\hat{\alpha}_i$ of each insurance group (main specification). The markers show each group's median fitted mean. (b) Observed and predicted shares of zero visits by insurance group.

7.4. Estimates

Table 8 reports the estimates of both equations. The clinical mean coefficients and their standard errors (SEs) are almost unchanged when the dispersion equation is added (for example, the SE of the age coefficient is 0.007 in both models), as predicted in Section 4.5. The standard errors of the insurance mean coefficients increase from 0.054 to 0.088 for public-only insurance and from 0.052 to 0.087 for any private insurance. This is not an effect of correlation between x_i and z_i : The weights

$$\alpha_i \mu_i / (\alpha_i + \mu_i)$$

in (4.7) are smaller for the uninsured once their α_i is allowed to be small.

Table 8. Estimates of the mean and dispersion equations (standard errors from the observed information). A positive $\hat{\eta}_j$ means a larger α_i , that is, less extra-Poisson variation at a given mean. Reference categories: poor income; uninsured.

Covariate	Mean equation, $\hat{\beta}$ (SE)		Dispersion equation, $\hat{\eta}$ (SE)		
	NB, const. α (mean $x + z$)	VD-NB (main) (mean $x + z$)	VD-NB (main) (mean $x + z$)	exp($\hat{\eta}$)	Restricted mean (mean x)
Intercept	−2.244 (0.082)	−2.223 (0.111)	−2.711 (0.104)	—	−3.298 (0.096)
Age (decades)	+0.144 (0.007)	+0.143 (0.007)	—	—	—
Self-rated physical health (worse)	+0.149 (0.014)	+0.154 (0.014)	—	—	—
Self-rated mental health (worse)	+0.146 (0.013)	+0.135 (0.013)	—	—	—
Hypertension	+0.161 (0.028)	+0.142 (0.028)	—	—	—
Coronary heart disease	+0.307 (0.051)	+0.293 (0.050)	—	—	—
Diabetes	+0.117 (0.035)	+0.101 (0.035)	—	—	—
Asthma	+0.337 (0.033)	+0.325 (0.032)	—	—	—
Female	+0.426 (0.023)	+0.411 (0.023)	+0.215 (0.027)	1.24	+0.236 (0.027)
Education (years)	+0.086 (0.004)	+0.091 (0.005)	+0.044 (0.005)	1.05	+0.048 (0.005)
Near-poor income	+0.055 (0.063)	+0.074 (0.068)	+0.084 (0.075)	1.09	+0.103 (0.073)
Low income	+0.033 (0.046)	+0.005 (0.049)	+0.069 (0.055)	1.07	+0.083 (0.053)
Middle income	+0.083 (0.040)	+0.057 (0.043)	+0.216 (0.048)	1.24	+0.224 (0.046)
High income	+0.333 (0.041)	+0.291 (0.043)	+0.305 (0.048)	1.36	+0.335 (0.047)
Public-only insurance	+1.379 (0.054)	+1.354 (0.088)	+1.408 (0.081)	4.09	+1.856 (0.073)
Any private insurance	+1.311 (0.052)	+1.286 (0.087)	+1.238 (0.081)	3.45	+1.667 (0.073)
$\hat{\alpha}$ (constant model)	0.593				
Log-likelihood	−42,615.8	−42,328.8			−42,868.2

Placing z_i in the mean reduces the insurance dispersion coefficients from 1.86 and 1.67 to 1.41 and 1.24, that is, from multiplicative effects on α_i of 6.4 and 5.3 to 4.1 and 3.4. The income and education coefficients change little. Female sex, education, and middle and high income are associated with larger α_i , that is, with less extra-Poisson variation at a given mean. Figure 6 compares these effects across specifications. The following subsections examine whether they are robust.

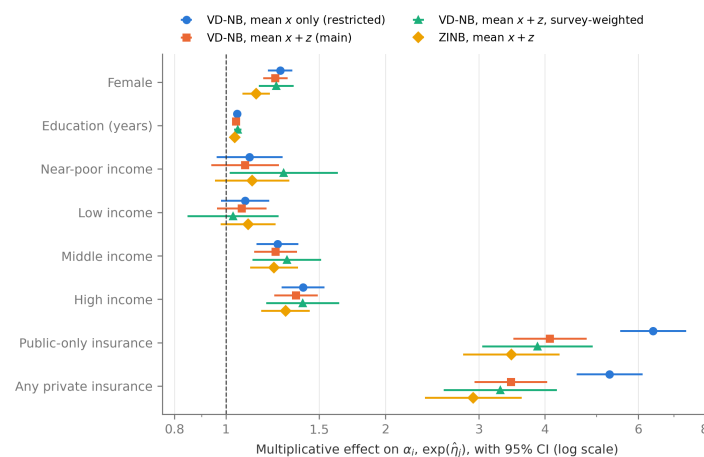


Figure 6. Dispersion effects $\exp(\hat{\eta}_j)$ with 95% confidence intervals under four specifications. The survey-weighted intervals use design-based (linearization) standard errors.

7.5. Sensitivity to the mean specification

Table 9 moves insurance and income into the mean one block at a time. Moving insurance accounts for almost all of the change in the test statistics and in the insurance dispersion coefficients. The score statistic falls from 1,109.5 to 704.9, and $\hat{\eta}$ for public-only insurance falls from 1.86 to 1.45. Moving income changes very little. The female, education, and income dispersion effects are stable across all five specifications.

Table 9. Sensitivity of the dispersion equation to the mean specification. The dispersion covariates are z_i in every column. Entries are $\hat{\eta}$ (standard error).

	x only	x + insurance	x + income	x + insurance + income	x + all z
Score (expected info.)	1,109.5	704.9	1,004.7	690.0	660.3
Score (robust)	789.1	230.5	663.4	210.1	164.4
Likelihood ratio	1,259.1	655.4	1,161.3	634.5	573.8
BIC	85,899.6	85,719.9	85,708.1	85,579.0	84,897.6
Female	+0.236 (0.027)	+0.232 (0.027)	+0.237 (0.027)	+0.230 (0.027)	+0.215 (0.027)
Education (years)	+0.048 (0.005)	+0.045 (0.005)	+0.045 (0.005)	+0.043 (0.005)	+0.044 (0.005)
Near-poor income	+0.103 (0.073)	+0.087 (0.074)	+0.101 (0.074)	+0.108 (0.074)	+0.084 (0.075)
Low income	+0.083 (0.053)	+0.055 (0.054)	+0.083 (0.054)	+0.069 (0.054)	+0.069 (0.055)
Middle income	+0.224 (0.046)	+0.185 (0.047)	+0.211 (0.046)	+0.194 (0.047)	+0.216 (0.048)
High income	+0.335 (0.047)	+0.315 (0.048)	+0.327 (0.047)	+0.320 (0.047)	+0.305 (0.048)
Public-only insurance	+1.856 (0.073)	+1.452 (0.079)	+1.828 (0.074)	+1.455 (0.080)	+1.408 (0.081)
Any private insurance	+1.667 (0.073)	+1.270 (0.079)	+1.648 (0.074)	+1.280 (0.079)	+1.238 (0.081)

7.6. Excess zeros

Under NB2, the probability of a zero at a given mean is $\{\alpha/(\alpha+\mu)\}^\alpha$, which increases as α decreases. A small α_i is therefore the only way the VD-NB2 model can reproduce the zero mass of the uninsured. Figure 5(b) shows that the constant-dispersion NB with z_i in the mean predicts 54.7% zeros among the uninsured, while the VD-NB model predicts 73.4%, close to the observed 72.9%. To separate the decision to use care from the amount of use, hurdle and zero-inflated NB2 (ZINB) models were fitted, with x_i and z_i in both the count and the zero components (Table 10).

Table 10. Zero-modified models, alternative variance functions, and smooth mean effects (k : number of parameters or effective degrees of freedom). In the hurdle row with insurance, the dispersion of uninsured positive counts lies on the boundary $\alpha \rightarrow 0$.

Model (mean: $x + z$)	k	Log-lik.	AIC	BIC	LR for z in dispersion (df)
<i>NB2 (Section 7)</i>					
NB2, constant α	17	-42,615.8	85,265.5	85,394.7	
NB2, α_i depends on z	25	-42,328.8	84,707.7	84,897.6	573.8 (8)
<i>Zero-modified NB2</i>					
Hurdle NB2, constant α	33	-41,974.1	84,014.2	84,264.9	
Hurdle NB2, α_i depends on z	41	-41,874.2	83,830.4	84,142.0	199.8 (8)
Hurdle NB2, α_i depends on z excl. insurance	39	-41,919.8	83,917.5	84,213.8	108.7 (6)
ZINB2, constant α	33	-42,279.3	84,624.7	84,875.4	
ZINB2, α_i depends on z	41	-42,158.6	84,399.1	84,710.6	241.5 (8)
<i>Alternative variance functions (GAMLSS)</i>					
NB1 (NBII), constant σ	17	-42,050.7	84,135.4	84,264.6	
NB1 (NBII), σ_i depends on z	25	-42,004.0	84,058.0	84,247.9	93.4 (8)
NB-P (NBF), constant σ	18	-42,032.1	84,100.2	84,236.9	
NB-P (NBF), σ_i depends on z	26	-42,001.2	84,054.5	84,252.0	61.7 (8)
<i>Smooth mean (GAMLSS, P-splines in age and education)</i>					
NB2, constant α	19.3	-42,587.6	85,213.8	85,360.4	
NB2, α_i depends on z	27.3	-42,307.2	84,669.1	84,876.6	560.7 (8)

Vuong tests [32] favor both the hurdle model ($V = 14.8$) and the ZINB model ($V = 10.3$) over the corresponding NB model, and so do AIC and BIC. Because the Vuong test is known to be unreliable for detecting zero inflation [33], the comparison rests mainly on the information criteria. Varying dispersion remains significant within both zero-modified models, with LR statistics of 241.5 for ZINB and 199.8 for the count component of the hurdle model. In the hurdle model, however, the dispersion of uninsured users with at least one visit reaches the boundary $\alpha \rightarrow 0$, where the truncated NB becomes a logarithmic-series distribution. The 299 uninsured users have a median of 2 visits, but some have up to 100, so their positive counts are extremely skewed. Without insurance in the hurdle dispersion equation, the remaining dispersion effects give an LR statistic of 108.7 on 6 degrees of freedom. The

estimates are 0.35 (SE 0.07) for female sex, 0.10 (0.02) for education, 0.20 (0.14) for middle income, and 0.27 (0.13) for high income. Their signs agree with those from VD-NB, but the female and education effects are larger, and the middle-income effect is no longer significant.

7.7. Variance function and comparison with GAMLSS

All models were also fitted with the `gamlss` package for R, version 5.5.0 [7]. The NBI family with $\log \sigma_i = -\log \alpha_i$ reproduces the VD-NB estimates exactly: The σ coefficients equal $-\hat{\eta}$ to three decimals, with identical standard errors, and the LR statistic is 573.8. Replacing the linear terms for age and education in the mean by penalized splines (P-splines) leaves the dispersion effects essentially unchanged (LR 560.7).

The variance function matters much more. NB1 (NBII) and NB-P (NBF, variance $\mu + \sigma\mu^p$) models with constant σ fit far better than constant-dispersion NB2. Their log-likelihoods are $-42,050.7$ and $-42,032.1$, against $-42,615.8$ for NB2, and the estimated power is $\hat{P} = 1.21$, much closer to NB1 than to NB2. Within these families, dispersion heterogeneity is much weaker. The LR statistics for σ_i depending on z_i are 93.4 (NB1) and 61.7 (NB-P) on 8 degrees of freedom. AIC still prefers varying dispersion, but BIC does not in the NB-P family. In the NB-P model, the insurance effects vanish: The coefficients are 0.06 (SE 0.13) for public-only and -0.04 (0.13) for private insurance. Middle and high income remain associated with less dispersion, with coefficients of -0.28 (0.06) and -0.21 (0.06) on the log σ scale, while the substantive direction of the female and education associations reverses under NB-P: Their coefficients are positive on the log σ scale, implying more rather than less dispersion. By Section 3.5, this pattern is expected when part of the NB2 dispersion heterogeneity reflects a variance function that is closer to linear than to quadratic in μ . It matters most for strong predictors of μ_i , such as the insurance indicators.

7.8. Choice of link

Table 11 illustrates Theorem 1(c). The expected-information score statistic equals 660.3 under all five links. The observed-information quadratic form ranges from -643.1 under the identity link, where the observed information is not positive definite and the quantity is not a valid score statistic, to 1,399.8 under the square-root link. The observed-information version should therefore not be used with links other than the logarithm. The full-model fits are similar under the log, log- ϕ , and identity- ϕ links (LR 573.1–573.8). The identity link on α fits worst but gives nearly the same ordering of individuals: The correlation of its fitted $\log \alpha_i$ with the log-link fit is 0.955.

Table 11. Sensitivity to the dispersion link (main specification).

Link $g(\cdot)$	Score (expected)	Score (observed)	LR	AIC	$\text{corr}(\log \hat{\alpha}_i, \log \hat{\alpha}_i^{\log})$
$\log \alpha_i$ (default)	660.3	602.9	573.8	84,707.7	1.000
$\sqrt{\alpha_i}$	660.3	1,399.8	550.5	84,731.0	0.990
α_i (identity)	660.3	-643.1	484.5	84,797.0	0.955
$\phi_i = 1/\alpha_i$ (identity)	660.3	412.0	573.1	84,708.5	0.992
$\log \phi_i = -\log \alpha_i$	660.3	602.9	573.8	84,707.7	1.000

7.9. Survey design

The analyses above treat the sample as independent observations. To account for the MEPS design, Table 12 reports weighted pseudo-maximum likelihood estimates with design-based (linearization) standard errors [34, 35]. A design-based score test was also computed from the linearization variance of the efficient score, a first-order Rao-Scott-type correction [36]. Weighting changes most dispersion estimates only slightly. The exception is near-poor income, whose coefficient rises from 0.084 to 0.251, although it remains imprecise (design SE 0.120). The design effects of the dispersion coefficients range from 1.8 to 2.6, and up to 4.5 for the mean coefficients. The design-based score statistic is 113.5 on 8 degrees of freedom, and the design-adjusted Wald statistic is $F_{8,150} = 30.6$. Both are much smaller than the model-based statistics, but both remain decisive. Inference about the U.S. adult population should be based on these design-based results.

Table 12. Survey-weighted estimates of the dispersion equation (main specification) with design-based standard errors. The design effect is the ratio of the design-based to the model-based variance.

Dispersion covariate	Unweighted $\hat{\eta}$ (model SE)	Weighted $\hat{\eta}$ (model SE)	Weighted (design SE)	Design effect
Female	+0.215 (0.027)	+0.219 (0.028)	(0.038)	1.94
Education (years)	+0.044 (0.005)	+0.051 (0.006)	(0.008)	2.09
Near-poor income	+0.084 (0.075)	+0.251 (0.090)	(0.120)	1.76
Low income	+0.069 (0.055)	+0.031 (0.063)	(0.101)	2.61
Middle income	+0.216 (0.048)	+0.265 (0.055)	(0.076)	1.91
High income	+0.305 (0.048)	+0.334 (0.055)	(0.081)	2.16
Public-only insurance	+1.408 (0.081)	+1.353 (0.083)	(0.122)	2.19
Any private insurance	+1.238 (0.081)	+1.193 (0.081)	(0.125)	2.40

7.10. Out-of-sample predictive performance

The final step of the workflow asks whether varying dispersion improves prediction. The models were compared by 10-fold cross-validation, using the same random folds for every model (Table 13). The main criteria are the mean held-out log score and the ranked probability score (RPS), both proper scoring rules for count distributions [37, 38]. The root mean squared error (RMSE), the relative RMSE (RRMSE, the RMSE divided by the sample mean; see, e.g., [39, 40]), the mean absolute error (MAE), and the coverage and width of central 90% prediction intervals are also reported.

Table 13. Ten-fold cross-validated predictive performance ($n = 14,732$). Higher log scores and lower RPS, RMSE, and MAE are better. RRMSE is the RMSE divided by the sample mean (7.97).

Model	Log score	RPS	RMSE	RRMSE (%)	MAE	PI ₉₀ cover.	PI ₉₀ width	Log score, uninsured
NB2, mean x	-2.9534	5.226	13.95	175.0	7.86	0.947	31.3	-1.654
VD-NB2, mean x , disp. z	-2.9113	5.127	13.97	175.3	8.23	0.952	34.4	-1.251
NB2, mean $x + z$	-2.8957	4.965	13.60	170.7	7.40	0.952	30.0	-1.307
VD-NB2, mean $x + z$, disp. z	-2.8767	4.949	13.57	170.3	7.36	0.951	29.1	-1.161
Hurdle NB2, const. α	-2.8528	4.921	13.53	169.9	7.34	0.950	30.0	-1.168
Hurdle NB2, disp. z excl. insurance	-2.8497	4.910	13.52	169.8	7.33	0.950	29.8	-1.165
Hurdle NB2, disp. z	-2.8465	4.910	13.52	169.7	7.33	0.949	29.6	-1.137

Varying dispersion improves the log score by 0.019 per observation (fold SE 0.003) under the main specification and by 0.042 (SE 0.002) under the restricted-mean specification, and it does so in all 10 folds. Point predictions barely change: RMSE is 13.57 versus 13.60, and RRMSE is 170–175% for every NB model, reflecting the high variability and heavy upper tail of the counts. Because dispersion primarily alters the predictive distribution rather than the conditional mean, squared-error criteria are relatively insensitive to its contribution. The hurdle NB models have the best log scores. Even with constant dispersion, the hurdle model improves on the VD-NB model by 0.024 (SE 0.002). Adding varying dispersion to the hurdle count component improves the log score further, but only slightly: by 0.006 (SE 0.001) with insurance in the dispersion equation and by 0.003 (SE 0.001) without it. Out of sample, the benefit of varying dispersion is therefore real but modest, and much of it can be obtained by modeling the zeros separately.

7.11. Summary of the application

Constant dispersion is rejected under every mean specification, zero-generating process, and variance function considered, and under design-based inference. The diagnostic steps nevertheless change what the rejection means. Under the restricted-mean specification, insurance appears to increase α about sixfold. This effect shrinks once insurance enters the mean. It largely reflects the zero mass of the uninsured and the quadratic NB2 variance function, and it disappears under NB-P. It should not be read as evidence that insurance makes utilization more regular. The most robust finding is a modest income gradient: Adults in higher-income households show less extra-Poisson variation in office visits at a given expected level. The high-income effect appears in the NB2, zero-modified, survey-weighted, and NB-P fits, although it is only marginally significant in the hurdle model. The middle-income effect is significant in all of these fits except the hurdle model, and under NB-P it is larger than the high-income effect. All associations are observational, and unmeasured morbidity, plan design, or provider supply could produce similar patterns.

8. Conclusions

The score test for covariate-varying dispersion in NB regression can be computed directly from the standard constant-dispersion fit. With expected information, the statistic is a quadratic form in a weighted covariance matrix of the dispersion covariates and is exactly invariant to the dispersion link. In the simulations, its empirical size stays between 0.03 and 0.06 from $n = 50$ upward. Because

the expected information is block diagonal between the mean and dispersion parameters, correlation between the two covariate sets does not weaken the test asymptotically, and power is governed by the weighted variance of the dispersion index.

A significant score test is the start of an analysis rather than its conclusion, as the MEPS application shows. The test detects departures from the constant-dispersion NB2 model in the direction of z_i , and such departures may arise not only from genuine dispersion heterogeneity but also from omitted mean effects, a separate zero-generating process, or a variance function that is not quadratic. Standard tools separate these explanations: Tests of the mean specification, hurdle and zero-inflated models, and NB1 or NB-P fits, the last three of which are available in GAMLSS. Because p -values carry little information in large samples, effect sizes and proper scoring rules are needed to complement them.

The reference distribution of the test is asymptotic; the expected-information version performed well in the designs studied, but designs with many dispersion covariates or very small expected counts may require bootstrap calibration. Because the test is formulated within the NB2 model, interpreting a rejection as dispersion heterogeneity relies on the checks in step 5 of the workflow. The MEPS findings rest on observational data, and the complete-case analysis could be biased if missingness is related to utilization. The direction of this bias cannot be determined from the observed data alone: Exclusion would tend to bias the estimated level and dispersion of utilization downward if excluded adults have systematically higher conditional use or heterogeneity, and upward in the opposite case. Only 2.6% of adults were excluded, which limits the number of affected observations but does not by itself eliminate selection bias.

The same construction extends to score tests for varying dispersion within the NB-P, hurdle, and zero-inflated models, although the block structure of Lemma 2 would need to be rederived for each. Penalized dispersion regression [4] could help select z_i from many candidates, at the cost of post-selection inference. The broader idea of modeling an ancillary distributional parameter with covariates may also be useful in mixture models, such as the mixture cure model for interval-censored competing risks of Wang et al. [41], whose incidence and latency specification is not negative binomial. Extending the present score construction to such models would require a new likelihood and information derivation.

Use of Generative-AI tools declaration

The author declares that AI tools were used in the revised version solely for language editing, improving clarity, checking the logical consistency of certain mathematical arguments, and reviewing code. All scientific content, results, and conclusions were developed and independently verified by the author, who takes full responsibility for the accuracy and content of the manuscript.

Acknowledgments

The author thanks the editor and four anonymous reviewers for their insightful comments, which substantially improved the theoretical framework, numerical protocol, and simulation study.

Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU266177].

Conflict of interest

The author declares no conflict of interest.

References

1. A. C. Cameron, P. K. Trivedi, *Regression analysis of count data*, Cambridge: Cambridge University Press, 2013. <https://doi.org/10.1017/CBO9781139013567>
2. J. M. Hilbe, *Negative binomial regression*, Cambridge: Cambridge University Press, 2011. <https://doi.org/10.1017/CBO9780511973420>
3. G. K. Smyth, Generalized linear models with varying dispersion, *J. R. Stat. Soc. B*, **51** (1989), 47–60. <https://doi.org/10.1111/j.2517-6161.1989.tb01747.x>
4. R. A. Rigby, D. M. Stasinopoulos, Generalized additive models for location, scale and shape, *J. R. Stat. Soc. C*, **54** (2005), 507–554. <https://doi.org/10.1111/j.1467-9876.2005.00510.x>
5. G. K. Smyth, A. P. Verbyla, Adjusted likelihood methods for modelling dispersion in generalized linear models, *Environmetrics*, **10** (1999), 695–709.
6. T. J. Cole, P. J. Green, Smoothing reference centile curves: the LMS method and penalized likelihood, *Stat. Med.*, **11** (1992), 1305–1319. <https://doi.org/10.1002/sim.4780111005>
7. D. M. Stasinopoulos, R. A. Rigby, Generalized additive models for location scale and shape (GAMLSS) in R, *J. Stat. Softw.*, **23** (2007), 1–46. <https://doi.org/10.18637/jss.v023.i07>
8. D. M. Stasinopoulos, R. A. Rigby, G. Z. Heller, V. Voudouris, F. De Bastiani, *Flexible regression and smoothing: using GAMLSS in R*, Boca Raton: Chapman and Hall/CRC, 2017. <https://doi.org/10.1201/b21973>
9. T. Kneib, A. Silbersdorff, B. Säfken, Rage against the mean—a review of distributional regression approaches, *Econom. Stat.*, **26** (2023), 99–123. <https://doi.org/10.1016/j.ecosta.2021.07.006>
10. A. C. Cameron, P. K. Trivedi, Econometric models based on count data: comparisons and applications of some estimators and tests, *J. Appl. Econom.*, **1** (1986), 29–53. <https://doi.org/10.1002/jae.3950010104>
11. W. Greene, Functional forms for the negative binomial model for count data, *Econ. Lett.*, **99** (2008), 585–590. <https://doi.org/10.1016/j.econlet.2007.10.015>
12. C. B. Dean, Testing for overdispersion in Poisson and binomial regression models, *J. Am. Stat. Assoc.*, **87** (1992), 451–457. <https://doi.org/10.1080/01621459.1992.10475225>
13. Z. Yang, J. W. Hardin, C. L. Addy, Score tests for overdispersion in zero-inflated Poisson mixed models, *Comput. Stat. Data Anal.*, **54** (2010), 1234–1246. <https://doi.org/10.1016/j.csda.2009.11.010>

14. C. R. Rao, Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation, *Math. Proc. Cambridge Philos. Soc.*, **44** (1948), 50–57. <https://doi.org/10.1017/S0305004100023987>
15. A. Buse, The likelihood ratio, Wald, and Lagrange multiplier tests: an expository note, *Am. Stat.*, **36** (1982), 153–157. <https://doi.org/10.1080/00031305.1982.10482817>
16. T. S. Breusch, A. R. Pagan, A simple test for heteroscedasticity and random coefficient variation, *Econometrica*, **47** (1979), 1287–1294. <https://doi.org/10.2307/1911963>
17. J. B. Ramsey, Tests for specification errors in classical linear least-squares regression analysis, *J. R. Stat. Soc. B*, **31** (1969), 350–371. <https://doi.org/10.1111/j.2517-6161.1969.tb00796.x>
18. H. White, Maximum likelihood estimation of misspecified models, *Econometrica*, **50** (1982), 1–25. <https://doi.org/10.2307/1912526>
19. D. D. Boos, On generalized score tests, *Am. Stat.*, **46** (1992), 327–333. <https://doi.org/10.1080/00031305.1992.10475921>
20. P. Deb, P. K. Trivedi, The structure of demand for health care: latent class versus two-part models, *J. Health Econ.*, **21** (2002), 601–625. [https://doi.org/10.1016/S0167-6296\(02\)00008-5](https://doi.org/10.1016/S0167-6296(02)00008-5)
21. J. Mullahy, Specification and testing of some modified count data models, *J. Econometrics*, **33** (1986), 341–365. [https://doi.org/10.1016/0304-4076\(86\)90002-3](https://doi.org/10.1016/0304-4076(86)90002-3)
22. J. Mullahy, Much ado about two: reconsidering retransformation and the two-part model in health econometrics, *J. Health Econ.*, **17** (1998), 247–281. [https://doi.org/10.1016/S0167-6296\(98\)00030-7](https://doi.org/10.1016/S0167-6296(98)00030-7)
23. D. Lambert, Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34** (1992), 1–14. <https://doi.org/10.2307/1269547>
24. A. M. Jones, Health econometrics, In: *Handbook of health economics*, Amsterdam: Elsevier, 2000, 265–344. [https://doi.org/10.1016/S1574-0064\(00\)80165-1](https://doi.org/10.1016/S1574-0064(00)80165-1)
25. E. R. Walker, J. R. Cummings, J. M. Hockenberry, B. G. Druss, Insurance status, use of mental health services, and unmet need for mental health care in the United States, *Psychiatr. Serv.*, **66** (2015), 578–584. <https://doi.org/10.1176/appi.ps.201400248>
26. B. D. Sommers, R. J. Blendon, E. J. Orav, A. M. Epstein, Changes in utilization and health among low-income adults after Medicaid expansion or expanded private insurance, *JAMA Intern. Med.*, **176** (2016), 1501–1509. <https://doi.org/10.1001/jamainternmed.2016.4419>
27. B. Efron, R. J. Tibshirani, *An introduction to the bootstrap*, New York: Chapman and Hall, 1993. <https://doi.org/10.1201/9780429246593>
28. R. Davidson, J. G. MacKinnon, The size distortion of bootstrap tests, *Econom. Theory*, **15** (1999), 361–376. <https://doi.org/10.1017/S0266466699153040>
29. J. V. Bradley, Robustness?, *Br. J. Math. Stat. Psychol.*, **31** (1978), 144–152. <https://doi.org/10.1111/j.2044-8317.1978.tb00581.x>
30. Agency for Healthcare Research and Quality, Medical Expenditure Panel Survey, 2023 full-year consolidated data file HC-251, 2025. Available from: https://meps.ahrq.gov/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-251.

31. M. Lin, H. C. Lucas, G. Shmueli, Too big to fail: large samples and the p-value problem, *Inf. Syst. Res.*, **24** (2013), 906–917. <https://doi.org/10.1287/isre.2013.0480>
32. Q. H. Vuong, Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica*, **57** (1989), 307–333. <https://doi.org/10.2307/1912557>
33. P. Wilson, The misuse of the Vuong test for non-nested models to test for zero-inflation, *Econ. Lett.*, **127** (2015), 51–53. <https://doi.org/10.1016/j.econlet.2014.12.029>
34. D. A. Binder, On the variances of asymptotically normal estimators from complex surveys, *Int. Stat. Rev.*, **51** (1983), 279–292. <https://doi.org/10.2307/1402588>
35. T. Lumley, Analysis of complex survey samples, *J. Stat. Softw.*, **9** (2004), 1–19. <https://doi.org/10.18637/jss.v009.i08>
36. J. N. K. Rao, A. J. Scott, On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data, *Ann. Stat.*, **12** (1984), 46–60. <https://doi.org/10.1214/aos/1176346391>
37. T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction, and estimation, *J. Am. Stat. Assoc.*, **102** (2007), 359–378. <https://doi.org/10.1198/016214506000001437>
38. C. Czado, T. Gneiting, L. Held, Predictive model assessment for count data, *Biometrics*, **65** (2009), 1254–1261. <https://doi.org/10.1111/j.1541-0420.2009.01191.x>
39. X. Xu, Y. Zhang, Neural network predictions of the high-frequency CSI300 first distant futures trading volume, *Financ. Mark. Portf. Manag.*, **37** (2023), 191–207. <https://doi.org/10.1007/s11408-022-00421-y>
40. B. Jin, X. Xu, Office real estate price index forecasts through Gaussian process regressions for ten major Chinese cities, *Adv. Comput. Intell.*, **4** (2024), 8. <https://doi.org/10.1007/s43674-024-00075-5>
41. Y. Wang, B. Zhou, C. Fu, X. Xu, Modeling cure fractions in interval-censored competing risks data: an AFT mixture approach, *Electron. Res. Arch.*, **34** (2026), 6573–6613. <https://doi.org/10.3934/era.2026288>
42. W. K. Newey, D. McFadden, Large sample estimation and hypothesis testing, In: *Handbook of econometrics*, Amsterdam: Elsevier, 1994, 2111–2245. [https://doi.org/10.1016/S1573-4412\(05\)80005-4](https://doi.org/10.1016/S1573-4412(05)80005-4)
43. A. W. van der Vaart, *Asymptotic statistics*, Cambridge: Cambridge University Press, 1998. <https://doi.org/10.1017/CBO9780511802256>

Appendix

A. Proof of Lemma 1

Conditioning on v_i and integrating against the $\text{Gamma}(\alpha_i, \alpha_i)$ density gives, for $y = 0, 1, \dots$,

$$\Pr(Y_i = y) = \int_0^\infty \frac{e^{-\mu_i v} (\mu_i v)^y}{y!} \frac{\alpha_i^{\alpha_i} v^{\alpha_i-1} e^{-\alpha_i v}}{\Gamma(\alpha_i)} dv = \frac{\mu_i^y \alpha_i^{\alpha_i}}{y! \Gamma(\alpha_i)} \int_0^\infty v^{y+\alpha_i-1} e^{-(\mu_i+\alpha_i)v} dv$$

$$= \frac{\mu_i^y \alpha_i^{\alpha_i}}{y! \Gamma(\alpha_i)} \cdot \frac{\Gamma(y + \alpha_i)}{(\mu_i + \alpha_i)^{y + \alpha_i}} = \frac{\Gamma(y + \alpha_i)}{\Gamma(\alpha_i) y!} \left(\frac{\alpha_i}{\alpha_i + \mu_i} \right)^{\alpha_i} \left(\frac{\mu_i}{\alpha_i + \mu_i} \right)^y,$$

which is (3.1). By the laws of total expectation and total variance,

$$\mathbb{E}(Y_i) = \mathbb{E}(\mu_i \nu_i) = \mu_i,$$

$$\text{Var}(Y_i) = \mathbb{E}\{\text{Var}(Y_i | \nu_i)\} + \text{Var}\{\mathbb{E}(Y_i | \nu_i)\} = \mu_i + \mu_i^2 \text{Var}(\nu_i) = \mu_i + \mu_i^2 / \alpha_i.$$

The argument does not use the value of α_i , so it applies unchanged when α_i depends on covariates. Taking $\alpha_i = \mu_i / \delta$ gives the NB1 variance $\mu_i(1 + \delta)$, and taking $\alpha_i = \mu_i^{2-P} / \sigma$ gives the NB-P variance $\mu_i + \sigma \mu_i^P$ (Section 3.5). $\square$

B. Derivatives, information, and proof of Lemma 2

General links. Let $\alpha_i = h(\omega_i)$ with $\omega_i = w_i^\top \gamma$ and $h = g^{-1}$. Writing ℓ_α and $\ell_{\alpha\alpha}$ for the derivatives in (4.3)–(4.5), and $\ell_{\alpha\alpha,i} = \mu_i(Y_i - \mu_i) / s_i^2$,

$$\begin{aligned} \frac{\partial \ell_i}{\partial \gamma} &= w_i \ell_{\alpha,i} h'(\omega_i), \\ \frac{\partial^2 \ell_i}{\partial \gamma \partial \gamma^\top} &= w_i w_i^\top \{ \ell_{\alpha\alpha,i} h'(\omega_i)^2 + \ell_{\alpha,i} h''(\omega_i) \}, \\ \frac{\partial^2 \ell_i}{\partial \beta \partial \gamma^\top} &= \tilde{x}_i w_i^\top \ell_{\alpha\alpha,i} h'(\omega_i), \\ \frac{\partial^2 \ell_i}{\partial \beta \partial \beta^\top} &= -\tilde{x}_i \tilde{x}_i^\top \frac{\alpha_i \mu_i (\alpha_i + Y_i)}{s_i^2}. \end{aligned}$$

For the log link, $h' = h'' = \alpha$, which gives (4.2)–(4.5). For the square-root link, $(h', h'') = (2\sqrt{\alpha}, 2)$. For the identity link on $\phi = 1/\alpha$, $(h', h'') = (-\alpha^2, 2\alpha^3)$. All analytic derivatives in the accompanying code were checked against central differences (maximum relative error below 10^{-7}). The expected information was checked against a Monte Carlo average of 2,000 observed Hessians (maximum relative error 0.13%).

Proof of Lemma 2.

(i) Since

$$\begin{aligned} \mathbb{E}(Y_i) &= \mu_i, \\ \mathbb{E}\left\{-\frac{\partial^2 \ell_i}{\partial a_i^2}\right\} &= \frac{\alpha_i \mu_i (\alpha_i + \mu_i)}{s_i^2} = \frac{\alpha_i \mu_i}{s_i}, \\ \mathbb{E}\{\ell_{\alpha\alpha,i}\} &= 0. \end{aligned}$$

This gives the β block and the zero off-diagonal block for any link.

(ii) By the information identity, $\mathbb{E}(\ell_{\alpha,i}) = 0$ and $\mathbb{E}(-\ell_{\alpha\alpha,i}) = \mathbb{E}(\ell_{\alpha,i}^2)$. Hence the γ block is

$$w_i w_i^\top h'(\omega_i)^2 \mathbb{E}(-\ell_{\alpha\alpha,i}),$$

and the term involving h'' has mean zero. Since

$$\begin{aligned}\psi'(x+1) &= \psi'(x) - x^{-2}, \\ \psi'(Y+\alpha) - \psi'(\alpha) &= -\sum_{k=0}^{Y-1} (\alpha+k)^{-2}.\end{aligned}$$

By Tonelli's theorem,

$$\mathbb{E}\{\psi'(Y+\alpha)\} - \psi'(\alpha) = -\sum_{k \geq 0} \Pr(Y > k)(\alpha+k)^{-2}.$$

Since $\mathbb{E}(\mu - Y) = 0$, it follows that

$$\mathbb{E}(-\ell_{\alpha\alpha}) = \sum_{k \geq 0} \pi_k / (\alpha+k)^2 - 1/\alpha + 1/(\alpha+\mu) = i_{\alpha\alpha}(\mu, \alpha).$$

With $h' = \alpha$, this gives (4.7).

(iii) Positivity: $i_{\alpha\alpha} = \text{Var}(\ell_{\alpha})$, and ℓ_{α} is a non-constant function of Y . Indeed,

$$\ell_{\alpha}(y+1) - \ell_{\alpha}(y) = \frac{1}{y+\alpha} - \frac{1}{s},$$

which is nonzero except at the single value $y = \mu$. Since Y takes every nonnegative integer value with positive probability when $\mu > 0$, $\text{Var}(\ell_{\alpha}) > 0$.

(iv) Limits:

$$\sum_k \pi_k / (\alpha+k)^2 \leq \alpha^{-2} \sum_k \pi_k = \mu/\alpha^2 \rightarrow 0$$

as $\mu \rightarrow 0$, and $-1/\alpha + 1/(\alpha+\mu) \rightarrow 0$. As $\mu \rightarrow \infty$, $\pi_k \rightarrow 1$ for each k and

$$\pi_k / (\alpha+k)^2 \leq (\alpha+k)^{-2},$$

which is summable. By dominated convergence, the sum tends to

$$\sum_k (\alpha+k)^{-2} = \psi'(\alpha),$$

and $1/(\alpha+\mu) \rightarrow 0$. □

C. Proof of Theorem 1

Write $\lambda = (\beta^{\top}, \eta_0)^{\top}$, $\theta = (\lambda^{\top}, \eta^{\top})^{\top}$, $\ell_n(\theta) = \sum_i \ell_i(\theta)$, and let $\theta_0 = (\lambda_0^{\top}, \mathbf{0}^{\top})^{\top}$ be the true value under H_0 . All constants C below are finite and do not depend on i or θ .

Step 1 (dominance). By (A2) and (A3), $0 < \underline{\mu} \leq \mu_i \leq \bar{\mu}$ and $0 < \underline{\alpha} \leq \alpha_i \leq \bar{\alpha}$ uniformly on Θ . Stirling's bounds give

$$|\log \Gamma(y+\alpha) - \log \Gamma(\alpha) - \log y!| \leq C\{1 + \log(1+y)\},$$

and

$$|y \log(\mu/s)| \leq Cy.$$

So $\sup_{\Theta} |\ell_i| \leq C(1 + Y_i)$. Since

$$|\psi(y + \alpha) - \psi(\alpha)| \leq C\{1 + \log(1 + y)\}$$

and ψ', ψ'' are bounded on $[\underline{\alpha}, \infty)$, every first, second, and third derivative of ℓ_i with respect to θ is bounded, uniformly on Θ , by $C(1 + Y_i)^2$ times a polynomial in the (bounded) covariates. The NB distribution has finite moments of all orders, uniformly on Θ , so all these envelopes are integrable.

Step 2 (consistency of the restricted MLE). Under H_0 , $\mathbb{E}\{\ell_i(\lambda, \mathbf{0})\}$ is uniquely maximized at λ_0 . This follows from the information inequality together with Proposition 1 applied to the constant-dispersion model. Continuity, compactness, and the envelope of Step 1 give a uniform law of large numbers, so $\hat{\lambda} \xrightarrow{P} \lambda_0$ [42, Theorem 2.1 and Lemma 2.4].

Step 3 (expansion of the score). The restricted MLE satisfies $U_{\lambda}(\hat{\theta}_0) = \mathbf{0}$ with probability approaching one. A mean-value expansion around θ_0 , with the Hessian evaluated between $\hat{\theta}_0$ and θ_0 , together with the uniform law of large numbers for second derivatives (Step 1), gives

$$\sqrt{n}(\hat{\lambda} - \lambda_0) = \mathcal{I}_{\lambda\lambda}^{-1} n^{-1/2} U_{\lambda}(\theta_0) + o_p(1)$$

and

$$n^{-1/2} U_{\eta}(\hat{\theta}_0) = n^{-1/2} \{U_{\eta}(\theta_0) - \mathcal{I}_{\eta\lambda} \mathcal{I}_{\lambda\lambda}^{-1} U_{\lambda}(\theta_0)\} + o_p(1),$$

where $\mathcal{I}$ denotes the per-observation Fisher information at θ_0 . The summands in braces are i.i.d. with mean zero and covariance $\mathcal{I}_{\eta\eta\cdot\lambda}$. By Lemma 2, $\mathcal{I}_{\eta\beta} = \mathbf{0}$, and so

$$\mathcal{I}_{\eta\eta\cdot\lambda} = \mathbb{E}(c_i z_i z_i^{\top}) - \mathbb{E}(c_i z_i) \mathbb{E}(c_i z_i^{\top}) / \mathbb{E}(c_i) = \mathcal{I}^*,$$

which is positive definite. The multivariate central limit theorem gives

$$n^{-1/2} U_{\eta}(\hat{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}^*).$$

Step 4 (consistency of the information estimators). The function $c(\mu, \alpha)$ is continuous: Each term of the series is continuous, and the series is dominated by $\sum_k (\underline{\alpha} + k)^{-2}$. Hence $n^{-1} \hat{\mathcal{I}}_{\eta\eta\cdot\lambda}$ in (4.8) is a continuous function of $\hat{\theta}_0$ and of sample averages of bounded functions, and it converges in probability to $\mathcal{I}^*$. The same holds for the observed-information Schur complement (by the uniform law of large numbers for the Hessian and the information identity) and for the robust estimator $n^{-1} \sum_i \hat{U}_i \hat{U}_i^{\top}$ (by the uniform law of large numbers for outer products, whose envelope is $C(1 + Y_i)^4$). Part (a) then follows from the continuous mapping theorem.

Step 5 (local alternatives). The model is differentiable in quadratic mean at θ_0 : The log-density is continuously differentiable, and its score has a continuous, nonsingular covariance matrix (Step 1 and Lemma 2). Hence the model is locally asymptotically normal, and the sequence P_{θ_n} , with $\theta_n = (\lambda_0^{\top}, h^{\top} / \sqrt{n})^{\top}$, is contiguous to P_{θ_0} [43, Theorem 7.2]. Under P_{θ_0} , the efficient score $n^{-1/2} \tilde{U}_{\eta}$ and the log-likelihood ratio $\log dP_{\theta_n} / dP_{\theta_0}$ are jointly asymptotically normal, with asymptotic covariance $\mathcal{I}^* h$. The covariance with the non-efficient part of the score vanishes by construction. Le Cam's third lemma [43, Example 6.7] gives

$$n^{-1/2} U_{\eta}(\hat{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathcal{I}^* h, \mathcal{I}^*)$$

under P_{θ_n} . By contiguity, the information estimators still converge to $\mathcal{I}^*$, and so

$$S \xrightarrow{d} \chi_q^2(h^\top \mathcal{I}^* h).$$

This is (4.10).

Step 6 (link invariance). The restricted MLE of α is the same under every link (invariance of maximum likelihood), and at $\eta = \mathbf{0}$,

$$h'(\omega_i) = h'(g(\hat{\alpha})) = 1/g'(\hat{\alpha})$$

is the same for every i . By Appendix B, the score for η under link g equals $\{\hat{\alpha}g'(\hat{\alpha})\}^{-1}$ times the log-link score. The expected information blocks for (η_0, η) equal $\{\hat{\alpha}g'(\hat{\alpha})\}^{-2}$ times their log-link values, and the β blocks are unchanged and orthogonal. The Schur complement therefore scales by $\{\hat{\alpha}g'(\hat{\alpha})\}^{-2}$, and S is unchanged. With observed information, the term $\ell_{\alpha,i}h''(\omega_i)$ does not scale in this way, so the observed-information statistic is only asymptotically invariant. $\square$

D. Additional simulation results

Table A1. Empirical size at the 5% level ($R = 1,000$ per cell; Monte Carlo SE about 0.007).

Design	Test	$n = 50$	$n = 100$	$n = 200$	$n = 300$	$n = 500$	$n = 1000$	$n = 2000$
Shared	Score, expected	0.041	0.032	0.047	0.038	0.059	0.054	0.048
	Score, observed	0.091	0.071	0.065	0.054	0.064	0.056	0.052
	Score, robust	0.023	0.028	0.042	0.040	0.058	0.055	0.048
	LR	0.098	0.055	0.058	0.049	0.060	0.061	0.050
	Wald	0.044	0.026	0.042	0.034	0.052	0.051	0.048
Distinct	Score, expected	0.029	0.048	0.045	0.049	0.040	0.058	0.046
	Score, observed	0.111	0.097	0.054	0.056	0.047	0.056	0.049
	Score, robust	0.025	0.042	0.039	0.049	0.045	0.055	0.053
	LR	0.099	0.071	0.052	0.054	0.049	0.056	0.048
	Wald	0.025	0.025	0.031	0.040	0.040	0.052	0.046
Distinct, rescaled	Score, expected	0.053	0.038	0.053	0.047	0.052	0.051	0.051
	Score, observed	0.115	0.074	0.066	0.056	0.057	0.056	0.053
	Score, robust	0.029	0.031	0.046	0.048	0.054	0.058	0.052
	LR	0.112	0.056	0.065	0.059	0.056	0.055	0.053
	Wald	0.032	0.014	0.027	0.040	0.046	0.049	0.051

Table A2. Empirical power at the 5% level ($R = 1,000$ per cell). Rates for the LR test at $n \leq 100$ should be read together with its over-rejection under the null (Table A1).

Signal	Design	Test	$n = 50$	$n = 100$	$n = 200$	$n = 300$	$n = 500$	$n = 1000$	$n = 2000$
Weak	Shared	Score, expected	0.037	0.067	0.110	0.129	0.169	0.343	0.620
		LR	0.110	0.110	0.130	0.146	0.183	0.347	0.633
		Wald	0.059	0.055	0.105	0.130	0.171	0.352	0.634
	Distinct	Score, expected	0.040	0.056	0.075	0.076	0.104	0.178	0.311
		LR	0.101	0.088	0.089	0.085	0.114	0.183	0.315
		Wald	0.029	0.023	0.052	0.059	0.101	0.176	0.309
	Distinct, rescaled	Score, expected	0.043	0.066	0.084	0.121	0.194	0.348	0.631
		LR	0.111	0.093	0.098	0.143	0.203	0.357	0.634
		Wald	0.024	0.041	0.061	0.111	0.177	0.342	0.630
Moderate	Shared	Score, expected	0.088	0.213	0.463	0.654	0.873	0.993	1.000
		LR	0.200	0.290	0.494	0.691	0.897	0.995	1.000
		Wald	0.080	0.178	0.445	0.653	0.889	0.994	1.000
	Distinct	Score, expected	0.074	0.130	0.251	0.378	0.581	0.852	0.991
		LR	0.153	0.164	0.268	0.399	0.597	0.862	0.991
		Wald	0.038	0.058	0.188	0.318	0.559	0.853	0.991
	Distinct, rescaled	Score, expected	0.127	0.250	0.460	0.669	0.880	0.999	1.000
		LR	0.218	0.320	0.512	0.705	0.899	0.999	1.000
		Wald	0.049	0.134	0.400	0.642	0.876	0.999	1.000
Strong	Shared	Score, expected	0.180	0.444	0.779	0.927	0.999	1.000	1.000
		LR	0.303	0.530	0.823	0.944	0.999	1.000	1.000
		Wald	0.119	0.382	0.787	0.939	0.998	1.000	1.000
	Distinct	Score, expected	0.101	0.209	0.450	0.614	0.857	0.995	1.000
		LR	0.205	0.274	0.485	0.653	0.871	0.995	1.000
		Wald	0.050	0.091	0.367	0.598	0.856	0.995	1.000
	Distinct, rescaled	Score, expected	0.230	0.463	0.836	0.953	0.998	1.000	1.000
		LR	0.360	0.553	0.882	0.963	0.999	1.000	1.000
		Wald	0.080	0.290	0.803	0.951	0.998	1.000	1.000

E. Implementation

(1) Fit the constant-dispersion NB model by damped Newton-Raphson with the analytic Hessian. When the Hessian is not negative definite, the step falls back to Fisher scoring, and a step-halving line search is used throughout.

(2) Compute $\hat{u}_i$ and $\hat{c}_i$. The series in (4.6) is summed until $\Pr(Y > k) < 10^{-12}$ and $k > \hat{\mu}_i$. This takes about 0.3 seconds for $n = 14,732$ on one CPU core.

(3) Form U_η and (4.8), compute S , and refer it to χ_q^2 . For small n , repeat with $B = 199$ (or more) parametric bootstrap samples drawn from the fitted null model.

(4) If the test rejects, fit (3.3) by the same Newton algorithm, starting from $(\hat{\theta}_0^\top, \mathbf{0}^\top)^\top$, and carry out the checks in step 5 of Figure 1.

The central-difference Hessian is useful only as a check. It agrees closely with the analytic observed information (Appendix B); in the MEPS data the resulting statistics differ by less than 0.1%, and it is much more expensive to compute.



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)