Tatra Mt. Math. Publ. **26** (2003), 237–246

TATRA MOUNTAINS
*Mathematical Publications*

# IS LOGISTIC REGRESSION THE BEST FOR BINOMIAL RESPONSE?

Júlia Volaufová — Stephen M. Redmann, Jr. — Lynn R. LaMotte

ABSTRACT. In this paper we investigate a simple situation with one discrete explanatory variable, say $X$, with finite levels and the response variable being a vector of dichotomous responses, say $Y$, with replicated values at each level of $X$. The mean of $Y$ is modeled by a vector, say $p$ such that $E(Y) = p(X)$ and $\operatorname{var}(Y) = V$. The main question, whether the probability $p$ depends on levels of $X$ is addressed in several ways. First, considering a logistic regression analysis with both the Wald's test and the likelihood ratio test, second, looking at grouped data for each level of $X$, and using the angular transformation on proportions. The proportions are then modeled by linear model. Finally, the assumptions of discrete vector $Y$ are ignored and a weighted least squares (WLS) method is used on $E(Y) = p(X) = X'\gamma$ and $\operatorname{var}(Y) = \sigma^2 I$. In the latter case the estimates of $p(x)$ are truncated into 0 whenever the resulting probability estimate is negative. For estimation of $p$ we look at the total mean squared error with respect to the true value used in simulation. The more important part is the hypothesis of no dependence of $p$ on $X$, and here we compare the standard tests used within each method with respect to their size and power. The focus is mainly on the tails for small values of simulated probabilities and actual small sample sizes. We propose that the WLS yields equally good results as the logistic regression approach and the angularly transformed response of proportions.

## 1. Introduction and preliminaries

Common practice in statistical data analysis of dichotomous response and/or proportions in terms of one or more explanatory variables is to consider a generalized linear model approach for the analysis, commonly with a logit link function. Another standard way to analyze data in such situations is, particularly when dealing with proportions, to transform the response by the angular transformation in order to normalize the data and stabilize the variance, and then to proceed with a classic regression or linear model approach. In this paper, we attempt to address a simple question: what harm is done if we ignore the

2000 Mathematics Subject Classification: 62J05, 62F10.
Keywords: logistic regression, angular transformation, weighted least squares, Wald's test, likelihood ratio test.

fact that the response does not follow classic continuous variable distributional assumptions?

Let's consider the following situation. Let the response variable be a vector of independent dichotomous responses, say $Y$, with replicated values at each level $x$ of the explanatory variable $X$. We assume that the mean of $Y$ is modeled by a vector, say $p$, that is a function of the explanatory variable $X$.

More formally, let $y_1, y_2, \dots, y_n$ be binomial $\mathrm{Bin}(m_i, p_i)$ variables observed on $n$ sampling units. Let $x_1, x_2, \dots, x_n$ denote the corresponding values of the explanatory variable (not necessarily distinct). Then we have:

$$E(Y_i) = m_i p(x_i) \quad \text{and} \quad \mathrm{var}(Y_i) = m_i p(x_i)\big(1 - p(x_i)\big), \qquad i = 1, 2, \dots, n.$$

The main question, whether the probability $p$ depends on levels of $X$ can be addressed in several ways. Here we consider three approaches, the two mentioned above—the generalized linear model approach with a logit link and ordinary least squares (OLS) regression after the angular transformation applied to proportions —and third, ignoring the distributional assumptions, OLS regression applied to the response with respect to the explanatory variable, truncating estimates of $p(X)$ where necessary. Using simulated data we will compare the statistical properties of estimators of $p$ and also the size and power of test statistics testing the association between $p$ and the explanatory variable $X$.

### 1.1 Logistic regression

The generalized linear model approach with the logit link is widely covered in the literature: see, e.g., [2], and [1]. Here we consider the model defined by

$$\mathrm{logit}\big(p(X)\big) = \ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X.$$

The likelihood function $L$ is the product of $\binom{m_i}{y_i} p(x_i)^{y_i} \big(1 - p(x_i)\big)^{m_i - y_i}$, $i = 1, \dots, n$. Denote the maximum likelihood estimates of $\beta_0$ and $\beta_1$ by $\widehat{\beta_0}$ and $\widehat{\beta_1}$. The maximum likelihood estimator (MLE) of $p(X)$ is then

$$\widehat{p(X)} = \frac{\exp\big(\widehat{\beta_0} + \widehat{\beta_1} X\big)}{1 + \exp\big(\widehat{\beta_0} + \widehat{\beta_1} X\big)}.$$

We consider two statistics for testing $H_0\colon \beta_1 = 0$.

- The Wald statistic is

$$W = \frac{\widehat{\beta_1}}{\widehat{\mathrm{Stderr}}\,(\widehat{\beta_1})},$$

where $\widehat{\mathrm{Stderr}(\widehat{\beta_1})}$ is obtained from the inverse of the information matrix, evaluated at the MLEs.

- The logarithm of the likelihood ratio test statistic is

$$G = 2\ln\frac{L_1}{L_0},$$

where $L_1$ is the maximum of $L$ and $L_0$ is the maximum of $L$ restricted by $\beta_1 = 0$. Under $\mathrm{H}_0$ the distribution of $W$ is conventionally approximated by the standard normal distribution, and the distribution of $G$ by $\chi^2(1)$.

### 1.2 Angular transformation

In general, much less is known about the angular transformation of proportions and their use in analysis, although it is recommended and widely used. See, e.g., [3], p. 369, or [2], p. 137. It is defined by the following:

$$Z_i = 2\sqrt{m_i}\arcsin\left(\sqrt{\frac{Y_i}{m_i}}\right), \qquad i = 1, 2, \dots n.$$

It can be shown that the transformed random variables follow approximately normal distributions with a known identical variance equal to $1$, i.e.,

$$Z_i \approx N\big(E(Z_i), 1\big). \tag{1}$$

Using the Taylor series expansion, we can approximate the expected value of $Z_i$ by the first order or second order approximation. We will denote both the first and the second order approximation of $E(Z_i)$ by $\mu_{z_i}$. The first order approximation takes form

$$\mu_{z_i} \cong 2\sqrt{m_i}\arcsin\big(\sqrt{p_i}\big), \tag{2}$$

and the second order approximation is

$$\mu_{z_i} \cong 2\sqrt{m_i}\arcsin\big(\sqrt{p_i}\big) - \frac{1-2p_i}{4\sqrt{m_i p_i(1-p_i)}}. \tag{3}$$

Obviously, for $p_i = 0.5$ the two approximations coincide. The model for the approximate mean, $\mu_{z_i}$, is then

$$\mu_{z_i} = \alpha_0 + \alpha_1 x_i, \qquad i = 1, 2, \dots, n. \tag{4}$$

From (1) we can easily see that the estimates of $\alpha_0$ and $\alpha_1$ can be obtained from (4) by ordinary least squares estimation (OLS) applied to transformed observations $z_i$.

The OLS estimator of $p(x_i)$ is then a solution to equation (2) or equation (3) depending on whether we are interested in the first or second order approximation for the mean.

We will consider the test statistic for testing the hypothesis $H_0\colon \alpha_1 = 0$, formally expressed as

$$Q = \frac{\widehat{\alpha_1}}{\mathrm{Stderr}\,(\widehat{\alpha_1})}\,,$$

that under $H_0$ follows approximately the standard normal distribution.

### 1.3 Truncated weighted least squares approach

The mean of $\dfrac{E(Y_i)}{m_i}$ is $p(X_i)$ and its variance is $p(x_i)\big(1-p(x_i)\big)/m_i$. Consider a simple linear model in the additive form

$$Y_i/m_i = \gamma_0 + \gamma_1 X_i + \varepsilon_i\,, \qquad i = 1,2,\dots,n\,, \tag{5}$$

where the distribution of the random error term $\varepsilon_i$ is characterized by the first two moments

$$\varepsilon_i \sim \big(0, \sigma^2_{\varepsilon_i}\big)\,.$$

Let the solution to the OLS in model (5) be

$$\widetilde{p}_i = \widetilde{\gamma}_0 + \widetilde{\gamma}_1 x_i\,,$$

where $\widetilde{\gamma}_0$ and $\widetilde{\gamma}_1$ are results of the OLS. Since the true probabilities are between 0 and 1 we truncate the estimators as follows. If the OLS estimate of $p_i$ is $\le 0$ or $\ge 1$ then we put $\widetilde{p}_i = \delta$ or $\widetilde{p}_i = 1-\delta$, respectively, where $\delta$ is a small positive real number. In the second step of the iteration, we use $\widetilde{p}_i$ in the expression for the variance of $Y_i$ and hence we get

$$\widehat{p}_i = \widehat{\gamma}_0 + \widehat{\gamma}_1 x_i\,,$$

where $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$ are calculated from weighted LS with weights $m_i\big(\widetilde{p}_i(1-\widetilde{p}_i)\big)^{-1}$. Since we consider only the first two steps in the iteration process, the final estimates are truncated. We set

$$\widehat{p}_i = 0 \quad \text{or} \quad 1\,, \qquad \text{respectively,}$$

if the weighted estimate of $p_i$ is smaller than 0 or greater than 1.

The hypothesis $H_0\colon \gamma_1 = 0$ is tested by the test statistic

$$T = \frac{\widehat{\gamma}_1}{\widehat{\mathrm{Stderr}}\,(\widehat{\gamma}_1)}\,,$$

and compared to a critical value from a central $t$-distribution with $\sum\limits_{i=1}^{n} m_i - 2$ degrees of freedom.

## 2. Simulation

For the simulation, we generated observations from a linear logistic model with only one explanatory variable $X$. $X$ takes only discrete integer values in the range $-\frac{k-1}{2}, +\frac{k-1}{2}$ for $k = 3, 5, 7$. At each $x_i$, $m_i = m$ independent replicates of Bernoulli variable with $p(x_i)$ were generated with $m = 5, 10, 25$. The choice of parameters $\beta_0$ and $\beta_1$ was carefully designed as follows. Denote by $p_{\min} = p\left(X = -\frac{k-1}{2}\right)$ the probability at $X = -\frac{k-1}{2}$ and fix this value. We have used four different values of $p_{\min}$, 0.025, 0.1, 0.2, 0.3. Denote now by $p_{\max} = p\left(X = +\frac{k-1}{2}\right)$ and let its value be variable. Denote also $\delta_p = p_{\max} - p_{\min}$. With $p_{\min}$ fixed and $p_{\max}$ variable, we have a choice in determining $\delta_p$. We have chosen $\delta_p = 0$ to $0.2$ with step $0.01$. Then for $\beta_0$ and $\beta_1$ we get

$$\beta_0 = \frac{1}{2} \ln \frac{p_{\min}^2 + p_{\min}\delta_p}{(1 - p_{\min})(1 - p_{\min} - \delta_p)}$$

and

$$\beta_1 = \frac{1}{k-1} \ln \frac{(p_{\min} + \delta_p)(1 - p_{\min})}{(1 - p_{\min} - \delta_p)p_{\min}}.$$

As we can see, as we change $p_{\min}$ and $\delta_p$, both parameters $\beta_0$ and $\beta_1$ change. For each configuration of parameters we carried out 1000 simulations. The comparisons were on

- average loss,
- relative efficiency, and
- empirical size and power of a nominal 5 % test of $\beta_1 = 0$, which is equivalent to the hypothesis $\delta_p = 0$.

Average loss for each approach is defined as

$$R = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{1000} \sum_{j=1}^{1000} \left(\widehat{p_j(x_i)} - p(x_i)\right)^2,$$

where $\widehat{p_j(x_i)}$ is the estimated probability under the $j$th simulation and $p(x_i)$ is the true probability determined by the configuration of parameters. We concentrated more on the relative efficiency that has the form

$$E = \frac{R_{\text{logit}}}{R}.$$

The size and power for each test statistic are approximated in the most natural way—for each $p_{\min}$, for each $\delta_p$, and for each approach, the proportion

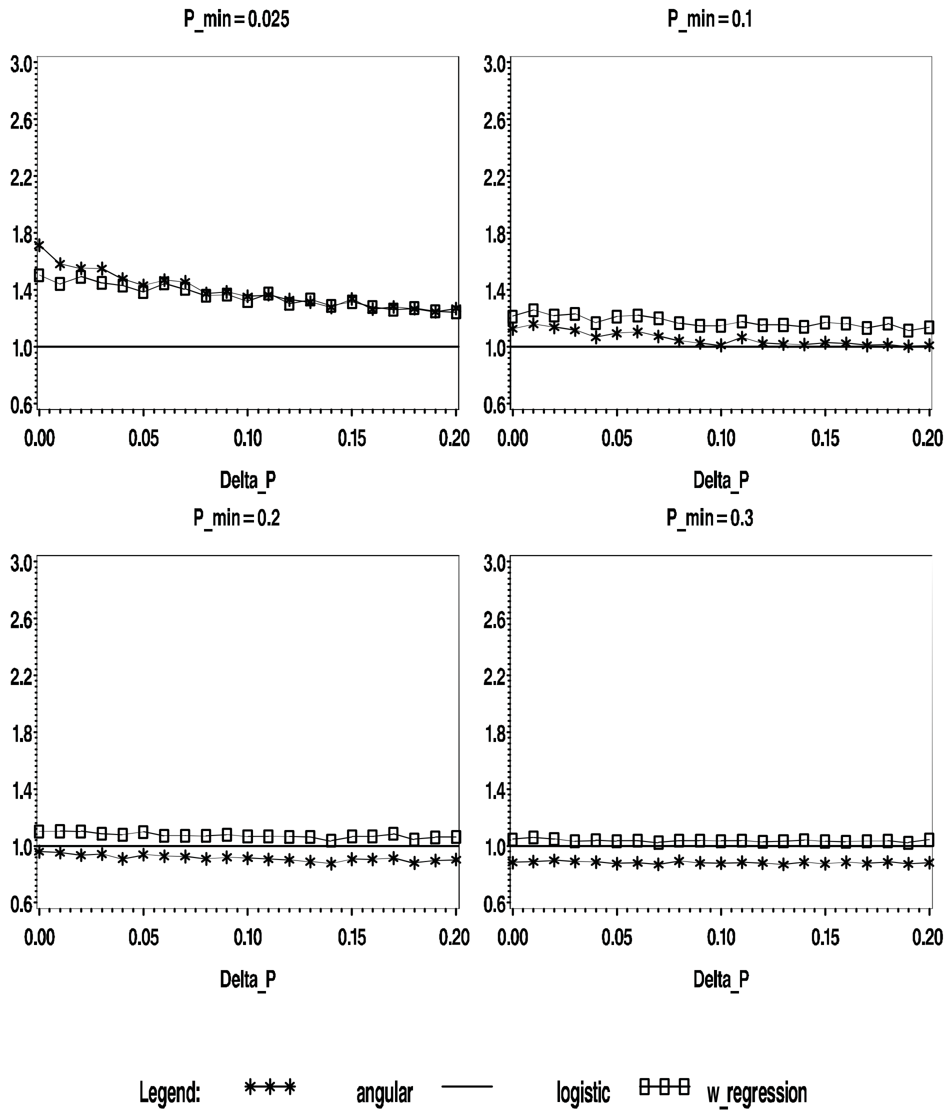


FIGURE 1. Relative efficiency of estimators of $p$. Number of levels of $X$, $k = 3$, number of Bernoulli replicates, $m = 5$.

of times the null hypothesis is rejected is the size ($\delta_p = 0$) or power ($\delta_p \neq 0$) of the test. Some of the results are illustrated by graphs (Figures 1–4).

## 3. Conclusion

With respect to relative efficiency, our results show that the logistic regression approach is not the most efficient method for estimating $p$. With small

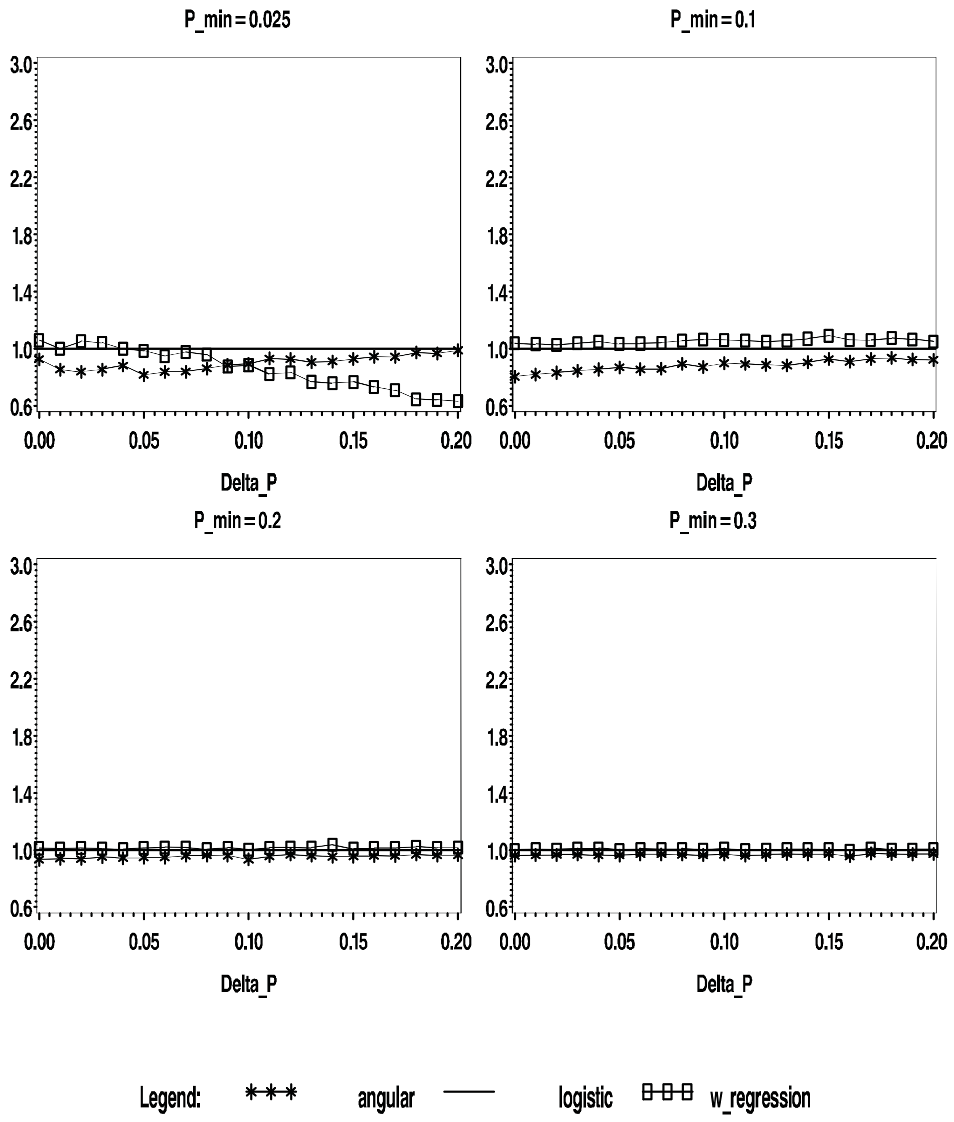


FIGURE 2. Relative efficiency of estimators of $p$. Number of levels of $X$, $k = 7$, number of Bernoulli replicates, $m = 25$.

number of levels of $X$, i.e., 3, 5 and 7, and for small $p_{\min}$, $p_{\min} = 0.025$, and for number of replicates that are equal to number of levels of $X$, in the whole range of $\delta$ the angular transformation approach is the most efficient method of estimating $p$ with respect to our criteria. As soon as $p_{\min}$ is larger, i.e., 0.1, 0.2, or 0.3, the weighted LS method, which ignores the fact that we are dealing with a dichotomous response, yields the most efficient result for all configurations of parameters and in the whole range of $\delta$. Regarding size and power of tests for $\delta_p = 0$, the situation is more complicated. Our results show that none of the

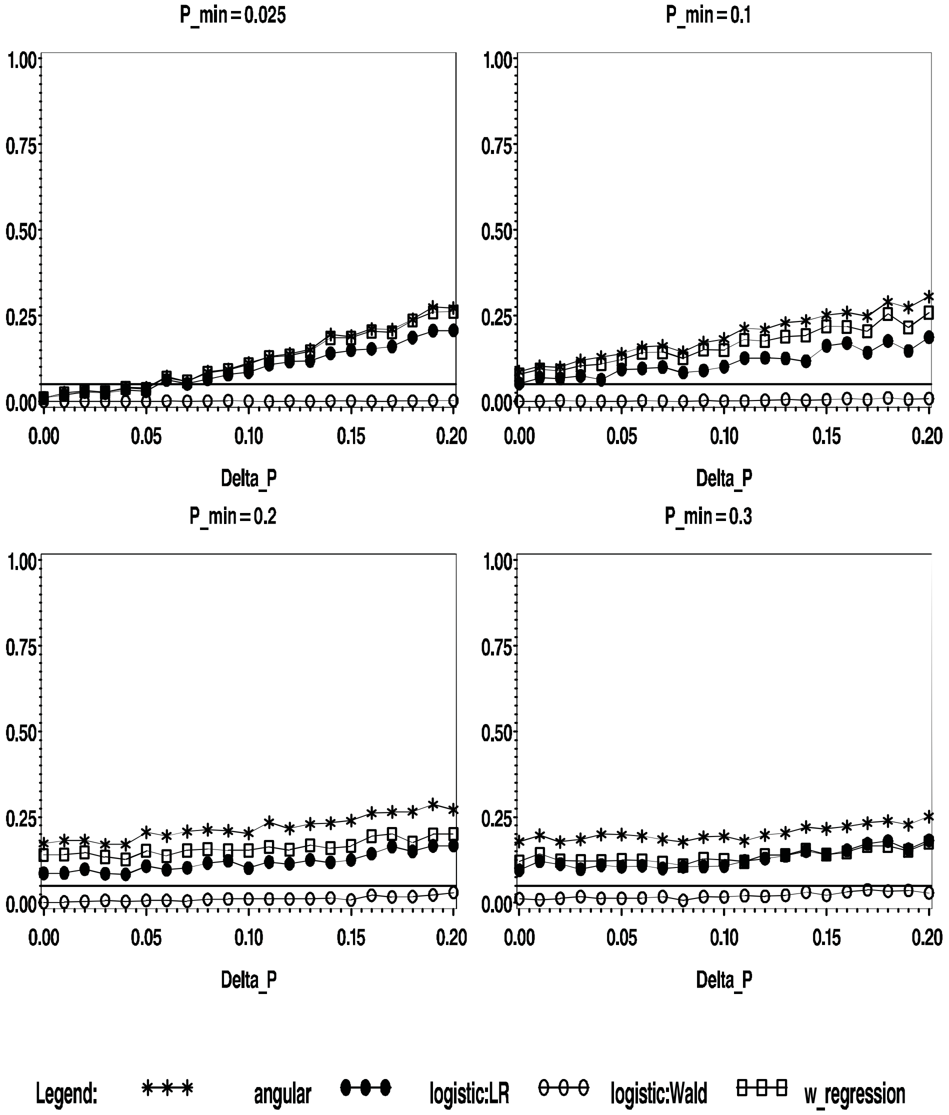


FIGURE 3. Size and power of tests for $\delta_p = 0$. Number of levels of $X$, $k = 3$, number of Bernoulli replicates, $m = 5$. The solid line indicates the 5 % significance level.

test statistics considered here, for small number of replicates at each level of $X$, reaches the nominal significance level $0.05$. As the number of replications increases, all tests seem to reach the size $0.05$. With the exception of $p_{\min} = 0.025$, in all cases the angular transformation approach results in the highest power, in most of the cases followed by weighted LS approach. It is worth mentioning that Wald's test under the logistic regression compared to the LR test statistic has very poor behavior with respect to size as well as power.

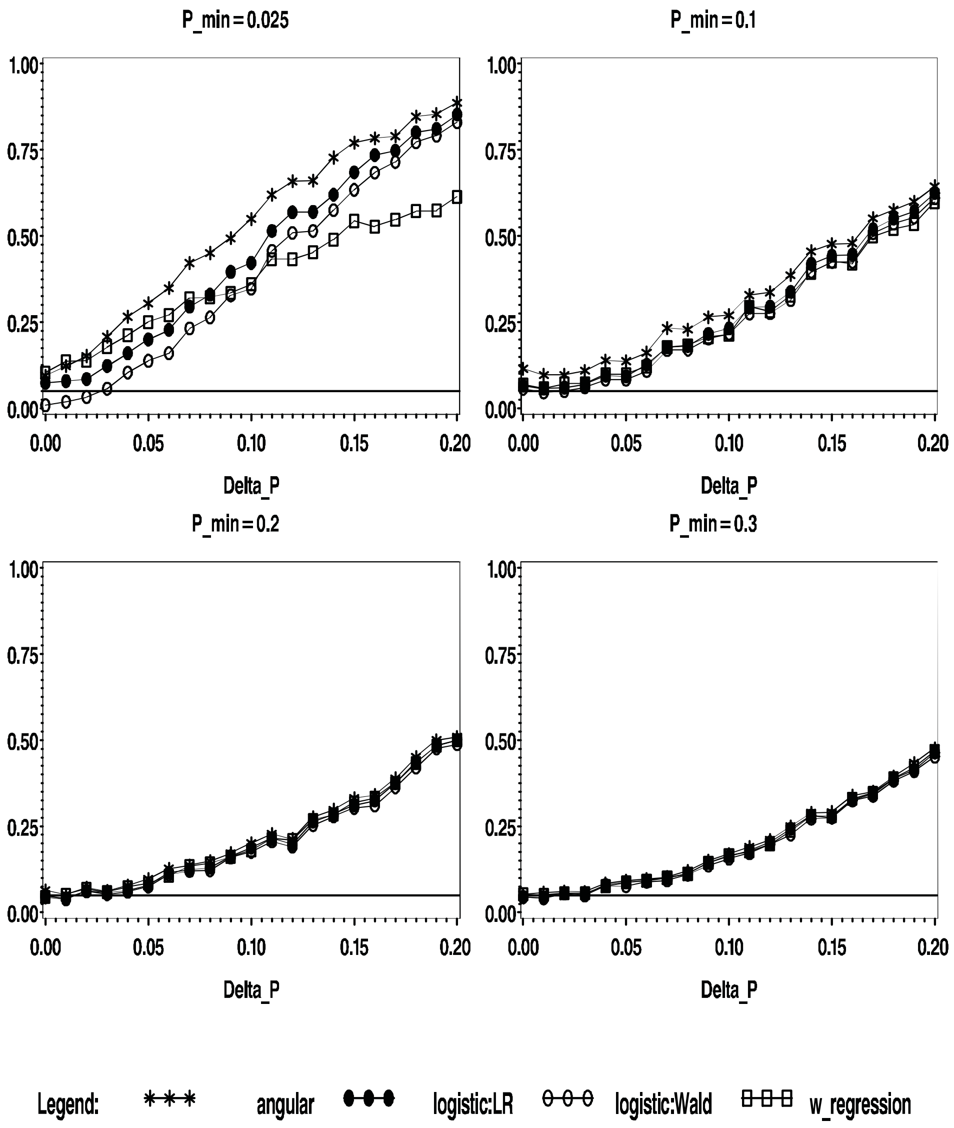


FIGURE 4. Size and power of tests for $\delta_p = 0$. Number of levels of $X$, $k = 7$, number of Bernoulli replicates, $m = 25$. The solid line indicates the 5 % significance level.

## REFERENCES

[1] HOSMER, D. W.—LEMESHOW, S.: *Applied Logistic Regression*, John Wiley & Sons, Inc. Second edition, New York, 2000.

[2] McCULLAGH, P.—NELDER, J. A.: *Generalized Linear Models*, Chapman & Hall/CRC. Second edition, Boca Raton, 1999.

[3] SEARLE, S. R.—CASELLA, G.—McCULLOCH, CH. E.: *Variance Components*, John Wiley & Sons, Inc., New York, 1992.

Received April 16, 2002

*Júlia Volaufová*
*Biostatistics and Data Management Core*
*Pennington Biomedical Research Center*
*LSU*
*6400 Perkins Road*
*Baton Rouge, LA 70808*
*U. S. A.*
*E-mail*: VolaufJ@mhs.pbrc.edu

*Stephen M. Redmann, Jr.*
*Biostatistics and Management Core*
*Pennington Biomedical Research Center*
*LSU*
*6400 Perkins Road*
*Baton Rouge, LA 70808*
*U. S. A.*

*Lynn R. LaMotte*
*Department of Experimental Statistics*
*LSU*
*161 Ag. Adm. Bldg.*
*Baton Rouge, LA 70803*
*U. S. A.*