MIMEOGRAPH SERIES #1633

Distribution-Free Tests in Binary Regression Models

E. Jacquelin Dietz

Department of Statistics, Box 5457, North Carolina State University

Raleigh, North Carolina 27650 U.S.A.

SUMMARY

This paper proposes the use of Wilcoxon rank sum, Wilcoxon signed rank, and sign statistics for testing certain hypotheses representing zero effects in the linear logistic regression model. A simulation study provides insight into the alternatives for which each test is most powerful. An example with real data is given for each test.

1. Introduction

Consider data $(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$, where $Y_1, Y_2, \cdots, Y_n$ are independent binary random variables with $P(Y_i = 1) = p_i$ and $P(Y_i = 0) = 1 - p_i$, $i = 1, 2, \cdots, n$. The x_i's are values of some explanatory variable thought to affect the p_i's. If $x_i = i$ for all i, the explanatory variable is simply the serial order of Y_i. Alternatively, x_i may be the dose level of a drug, and Y_i may correspond to death or survival.

Such data are frequently described by the linear logistic model

$$p_i = (1 + e^{-(\alpha + \beta x_i)})^{-1} \qquad (1)$$

or, equivalently,

$$\ell n(p_i/(1 - p_i)) = \alpha + \beta x_i ,$$

Key words: Linear logistic model; Rank sum test; Signed rank test; Sign test.

where α and β are unknown parameters. This model satisfies the constraint $0 \le p_i \le 1$ and yields simple sufficient statistics, $\sum_{i=1}^{n} Y_i$ and $\sum_{i=1}^{n} x_i Y_i$, similar to those for the normal-theory linear model (Cox, 1970).

In this paper, I discuss the use of these sufficient statistics and related rank statistics in testing certain hypotheses concerning α and β. The null hypotheses considered here are very general hypotheses that can be associated with models other than (1), for instance, with the probit model

$$p_i = \Phi(\alpha + \beta x_i) .$$

2. Hypothesis Tests

2.1 *Rank Sum Test*

Cox (1970) discusses a test of $H_0: \beta = 0$ versus $H_1: \beta \neq 0$ in (1), where α is a nuisance parameter. This test is based on the statistic $T = \sum_{i=1}^{n} x_i Y_i$ and is conditional on the observed value s of $S = \sum_{i=1}^{n} Y_i$. The statistic T is the total of a sample of size s chosen from the set $X = \{x_1, x_2, \cdots, x_n\}$. Under H_0, given s, all $\binom{n}{s}$ samples of size s drawn from X are equiprobable. A test of H_0 can be based on the resulting conditional distribution of T or, in some cases, on a normal approximation.

Cox points out that if $x_i = i$, $i = 1, 2, \cdots, n$, this conditional test is equivalent to the Wilcoxon rank sum test, with $Y_i = 1$ and $Y_i = 0$ identifying the usual two samples. Tables of the exact distribution of the rank sum statistic or the usual normal approximation can be used to carry out the test. If the x_i's are equally spaced, but not integer-valued, Cox's conditional test is equivalent to a rank sum test performed after ranking the x_i's. In fact, for any x_i's, a conditionally distribution-free test of H_0 can be obtained by replacing each x_i with its rank $R(x_i)$ and referring the statistic $\sum_{i=1}^{n} R(x_i) Y_i$ to the rank sum distribution. Ties among the x_i's can

be handled in the usual way. This test is equivalent to Kendall's τ test for association between the x_i's and Y_i's (Kendall, 1975, p. 165) and is consistent against alternatives with monotone increasing or decreasing p_i's, e.g., H_1.

2.2 *Signed Rank Test*

If $\alpha = 0$ in model (1), then T is the sufficient statistic for β. A test of $H_0^*: \alpha = \beta = 0$ versus $H_2: \alpha = 0, \beta \neq 0$ can be based on T. Under H_0^*, $p_i = .5$ for all i, and the 2^n possible sequences of zeros and ones are equally likely. There are thus 2^n equiprobable values of T corresponding to all possible subsets of X.

If the x_i's are equally spaced, the test based on T is equivalent to a test based on $\sum_{i=1}^{n} R(x_i)\, Y_i$. This statistic has the null distribution of the Wilcoxon signed rank statistic. For any x_i's, a distribution-free test of H_0^* can be based on this signed rank statistic and carried out using tables of its exact distribution or the usual normal approximation.

This test is consistent against alternatives $H_3: \alpha \neq 0, \beta = 0$. Thus, the signed rank statistic can be used to test H_0^* versus $H_4: \alpha \neq 0$ or $\beta \neq 0$.

Note that the rank sum and signed rank statistics are formally identical. However, for the former, s is fixed, while for the latter, s is the value of a random variable which can take on values $0, 1, \cdots, n$. Consequently, the two statistics can have very different null distributions and critical values (Lehmann, 1975, p. 125).

2.3 *Sign Test*

A second distribution-free test for H_0^* is based on the sign test statistic $S = \sum_{i=1}^{n} Y_i$. This test is consistent against alternatives H_3. For alternatives

with $\beta \neq 0$, results of Hoeffding (1956) are applicable. He shows that the usual one and two-sided tests for the constant probability of success in a binomial experiment can be used as tests for the average probability of success when the p_i's vary from trial to trial. Thus, the sign test can be used to test whether the average probability of success equals .5 in model (1). If $\beta \neq 0$ and $p = \sum_{i=1}^{n} p_i/n = .5$, the probability of rejecting H_0^* is less than the probability of rejecting H_0^* when H_0^* is true. Thus, the sign test is not unbiased for alternatives with $p = .5$. However, the power of the sign test for varying p_i's is at least as great as the power for constant $p = \sum_{i=1}^{n} p_i/n$ in cases where the latter power is sufficiently large (roughly when it exceeds .5). Since the sign test is consistent for any constant $p \neq .5$, it is also consistent for alternatives H_4 if n increases in such a way that $p = \sum_{i=1}^{n} p_i/n$ is constant and different from .5.

2.4 *Null and Alternative Hypotheses*

It is important to note that H_0 and H_0^* are very general hypotheses that need not be associated with the linear logistic model (1). H_0 specifies that $p_1 = p_2 = \cdots = p_n$, and H_0^* specifies that $p_1 = p_2 = \cdots = p_n = .5$. Model (1) is used here to motivate the choice of test statistics and to provide alternative distributions for the simulation study in § 3. The three distribution-free tests can be used for alternatives more general than those given by (1). In particular, no claim is made here that the linear logistic model provides a good description for the data sets used in § 4.

3. Simulation Study

3.1 *Methods*

The small sample power of the rank sum, signed rank, and sign tests for the linear logistic model was compared in a simulation study. Values of x_i

were taken to be equally spaced. Then, without loss of generality, x_i was set equal to i, $i = 1, 2, \cdots, n$, and different distributions obtained by varying α and β. Replacing a given α and β by $-\alpha$ and $-\beta$ does not affect the power of any of the tests; thus, only nonnegative values of β were considered. Positive values of β correspond to increasing p_i's. Five thousand samples were generated for each combination of the following parameter values:

$$n = 20, \quad \alpha = -2.5(.5).5, \quad \beta = .0(.05).25, \text{ and}$$

$$n = 40, \quad \alpha = -2.5(.5).0, \quad \beta = .0(.025).1.$$

Binary random variables $Y_1, Y_2, \cdots, Y_n$ were generated using the IMSL subroutine GGBN.

Each test was performed on each sample at a nominal significance level of .05, using the usual normal approximation for each statistic. Occasionally sequences of n zeros were obtained, making it impossible to carry out the rank sum test. Each such sample was counted as a rejection for the rank sum test, since a sequence of all zeros or all ones provides evidence against H_0^* (although not against H_0).

3.2 *Results and Discussion*

The results of the simulation study are shown in Table 1. Note that $p_i = .5$ when $x_i = -\alpha/\beta$. When $-\alpha/\beta = n/2$, the probabilities $p_1, p_2, \cdots, p_{n-1}$ are symmetric about .5. The cells of each subtable for which $-\alpha/\beta = n/2$ will be referred to as the diagonal of that subtable.

(INSERT TABLE 1 HERE)

The rank sum test is the most powerful of the three tests for values of α and β on and near the diagonal. For each value of α, the power increases

with β; for each value of β, the power decreases on either side of the diagonal.

Recall that sequences of n zeros were counted as rejections for the rank sum test. Such sequences are most frequent when $\beta = 0$, $|\alpha|$ is large, and n is small. If $H_0: \beta = 0$ is the null hypothesis of interest, H_0 should be accepted for such sequences. Using this decision rule, the following values of the power are obtained for $n = 20$ and $\beta = 0$:

α	power
-2.5	.02
-2.0	.03
-1.5	.04
-1.0	.04

The signed rank test is the most powerful of the three tests for values of α and β below the diagonal. For these alternatives, either all of the p_i's are greater than .5 or the p_i's cross .5 during the first half of the sequence. The lowest power for the signed rank test occurs immediately above the diagonal. For these alternatives, the p_i's cross .5 during the last half of the sequence.

These results point out a peculiarity of this test. Reversing the order of the sequence of zeros and ones changes the value of the signed rank statistic. This statistic is most sensitive to the Y_i's near the end of the sequence; if the p_i's near the end of the sequence are close to .5, the test may have low power. If the experimenter knows in advance that the alternatives of interest are of this type, a more powerful test can be obtained by ranking the x_i's from the end of the sequence. (See § 4.2.) For instance, when $n = 40$, $\alpha = -2$, and $\beta = .075$, the estimated power of the signed rank test is only .04. When x_i is set equal to $n - i + 1$,

$i = 1, 2, \cdots, n$, the estimated power increases to .65 for these parameter values.

The sign test is the most powerful of the three tests when $\beta = 0$ and when $-\alpha/\beta \geq n$. When $-\alpha/\beta \geq n$, $p_i \leq .5$ for all i. On the diagonal, the estimated power is less than .05, consistent with the results of Hoeffding (1956) discussed in §2.3. In each row and column of each subtable, the power increases moving away from the diagonal.

4. Examples

4.1 *Sign and Rank Sum Tests*

The tests based on the sign and rank sum statistics can be illustrated using data described in Real (1981) and Real, Ott, and Silverfine (1982). They study the effect of variability in nectar reward per flower on the foraging behavior of bumblebees, testing the theory that pollinators attempt not only to maximize expected energetic reward associated with foraging, but also to minimize uncertainty associated with that reward. The theory predicts that if two floral species offer the same expected reward, pollinators will prefer the species with less variability in reward.

In each study, a colony of bees was trained to forage on an artificial floral patch, consisting of 100 blue and 100 yellow cardboard squares randomly mixed and attached beneath a sheet of clear plexiglass. A shallow well in the plexiglass above each cardboard "flower" was filled with "nectar," a mixture of honey and water.

After filling each flower with a measured amount of nectar, an individual bee was observed foraging on the board. For a sequence of n visits to flowers, x_i is i, the serial order of the visit, for $i = 1, 2, \cdots, n$. Visits to blue and yellow flowers correspond to, say, $Y_i = 1$ and $Y_i = 0$, respectively.

Note that in this example, the Y_i's may not be mutually independent; however, independence can be incorporated into our null hypotheses. Such sequences were observed for different bees and different amounts of nectar per flower.

To test for color preference, each flower was filled with 2 μL of nectar. For a bee with no color preference, $p_i = .5$ for all i, corresponding to

$$H_0^*: \alpha = 0, \beta = 0, \text{ and } Y_1, Y_2, \cdots, Y_n \text{ independent.}$$

The alternative hypothesis of primary interest is $\beta = 0$, $\alpha \neq 0$, corresponding to a constant preference for one of the two colors. The sign test was seen in § 3.2 to be particularly powerful for such alternatives.

One bee in the study by Real *et al.* (1982) gave the sequence

yyyyyyyybbyyyyybbyybbyyyyybbybybbybyyyyy.

For this sequence, $n = 40$ and $\sum_{i=1}^{n} Y_i = 12$. The usual normal approximation for the distribution of the sign test statistic gives a normal deviate of -2.53, significant at $p = .0114$ (two-sided test). Thus, for this bee, there is evidence of a color preference, apparently for yellow flowers.

Other sequences were run to test the effects of variability in nectar reward. Real (1981) ran 15 trials in which every blue flower contained 2 μL of nectar. Every third yellow flower contained 6 μL of nectar; the remaining yellow flowers were empty. Thus, the blue and yellow flowers had the same expected reward, but differed in the variability of the reward.

If a bee is insensitive to variability in energetic reward, p_i should be constant (although not necessarily equal to .5) for all i, corresponding to

$$H_0: \beta = 0 \text{ and } Y_1, Y_2, \cdots, Y_n \text{ independent.}$$

If a bee is sensitive to variability in reward, it should learn that yellow flowers are variable and begin to avoid them. Thus, the probability of choosing a blue flower should increase over the trial, corresponding to alternatives $\beta > 0$. The appropriate test for H_0 is the rank sum test.

The first bee in this series gave the sequence

yyyyybbybbbyybybbbbbyybbbbbyybbbbb

(Real, 1981, p.25). For this sequence, $n = 34$, $\sum_{i=1}^{n} Y_i = 21$, and $\sum_{i=1}^{n} x_i Y_i = 432$. The normal approximation for the distribution of the rank sum statistic gives a normal deviate of 2.29, significant at $p = .0110$ (one-sided test). Thus, there is evidence of an increasing preference for blue flowers for this bee.

4.2 *Signed Rank Test*

The signed rank test can be illustrated using data described in Gottlieb (1981). In a series of previous studies (see references in Gottlieb, 1981), he demonstrated that the domestic mallard duck embryo must be exposed to its normal embryonic contact-contentment call in order to exhibit the species-typical preference for the normal species maternal call after hatching. Embryos that are muted and isolated from sib vocalizations show no preference for a recording of the normal maternal call over an artificially slowed recording of the call 48 hours after hatching. However, muted and isolated ducklings that are exposed to a recorded embryonic call for five minutes an hour during the last 48 hours of embryonic development show a strong preference for the normal maternal call 48 hours after hatching.

In Gottlieb (1981), muted and isolated ducklings were exposed to a recorded embryonic call for five minutes an hour for a period of 24 hours. The 24 hour period of stimulation began at different times relative to the time of

hatching for different ducklings. Each duckling was tested 48 hours after hatching for preference between the normal maternal call and a slowed maternal call. Table 2 shows the time of onset of stimulation in hours after hatching for each duckling and the call later preferred by that duckling. These ducklings are those in Experiments 1A and 1B of Gottlieb (1981) that exhibited a preference for one maternal call over the other according to the criterion described in that paper.

(INSERT TABLE 2 HERE)

For these data, x_i is the time of onset of stimulation for the i^{th} duckling, $i = 1, 2, \cdots, n$. Preference for the normal and slowed maternal calls can be associated with, say, $Y_i = 1$ and $Y_i = 0$, respectively. A question of interest in this study was whether 24 hours of stimulation was sufficient to result in preference for the normal maternal call. If not, then $p_i = .5$ for all i, corresponding to H_0^*. Considerations discussed in Gottlieb (1981) suggest that auditory stimulation might be most effective prenatally. If this is so, ducklings whose stimulation began early might be most likely to later prefer the normal call; those whose stimulation began later might not exhibit a preference.

Recall that the signed rank test had low power for such alternatives. However, a more powerful test can be obtained by ranking the x_i's from largest to smallest. The resulting test gives $n = 60$, $\sum_{i=1}^{n} R(x_i)Y_i = 1189.5$, and a normal deviate of 2.02, significant at $p = .0434$ (two-sided test). If the x_i's are ranked from smallest to largest, the signed rank test is not significant ($p = .2628$). The fact that the two signed rank tests give such different results suggests that the p_i's are not constant. The rank sum test is not significant for these data ($p = .3524$); however, the simulation study shows that the power of the rank sum test is relatively low for the type of alternative distribution anticipated here.

These signed rank tests suggest that early stimulation is more effective than later stimulation in producing preference for the normal maternal call 48 hours after hatching. Gottlieb (1981) concludes from additional experiments that postnatal stimulation is also effective, if ducklings are permitted a 48 hour "consolidation" period between stimulation and testing. In the preferred signed rank test, the ducklings are ranked (smallest to largest) according to the consolidation time received.

ACKNOWLEDGEMENTS

I wish to thank Leslie Real for permitting me to use his bumblebee data and Gilbert Gottlieb and Timothy Johnston for providing the duckling data.

REFERENCES

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

Gottlieb, G. (1981). Development of species identification in ducklings: VIII. Embryonic versus postnatal critical period for the maintenance of species-typical perception. *Journal of Comparative and Physiological Psychology* 95, 540-547.

Hoeffding, W. (1956). On the distribution of the number of successes in independent trials. *Annals of Mathematical Statistics* 27, 713-721.

Kendall, M. G. (1975). *Rank Correlation Methods*. London: Griffin.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

Real, L. A. (1981). Uncertainty and pollinator-plant interactions: the foraging behavior of bees and wasps on artificial flowers. *Ecology* 62: 20-26.

Real, L. A., Ott, J. and Silverfine, E. (1982). On the trade-off between the mean and the variance in foraging: effects of spatial distribution and color preference. *Ecology* 63,1617-1623.

TABLE 1

Estimated power of rank sum, signed rank, and sign tests

β	α						
	-2.5	-2.0	-1.5	-1.0	-.5	.0	.5
			Rank sum test, n = 20				
.00	.23	.11	.06	.05	.05	.04	.04
.05	.11	.08	.09	.09	.09*	.09	.08
.10	.15	.19	.21*	.23*	.22	.18	.15
.15	.37*	.41*	.42*	.40	.35	.28	.22
.20	.62*	.63*	.61*	.53	.44	.34	.28
.25	.80*	.77*	.72	.60	.48	.37	.34
			Signed rank test, n = 20				
.00	.98	.93	.76	.46	.15	.05	.16
.05	.87	.65	.31	.09	.05	.23*	.57*
.10	.52	.21	.06	.10	.33*	.67*	.89*
.15	.11	.04	.14	.41*	.73*	.92*	.98*
.20	.04	.18	.48	.77*	.94*	.99*	1.00*
.25	.23	.54	.82*	.95*	.99*	1.00*	1.00*
			Sign test, n = 20				
.00	1.00*	.97*	.85*	.55*	.16*	.04	.18*
.05	.97*	.84*	.50*	.16*	.04	.18	.54
.10	.79*	.45*	.13	.04	.17	.54	.85
.15	.35	.09	.02	.14	.47	.81	.95
.20	.06	.02	.12	.40	.75	.94	.99
.25	.01	.09	.34	.67	.89	.98	1.00

TABLE I continued

β	α						
	-2.5	-2.0	-1.5	-1.0	-.5	.0	.5
			Rank sum test, n = 40				
.000	.08	.05	.05	.05	.05	.05	
.025	.09	.10	.12	.14	.14*	.14	
.050	.28	.35	.40*	.43*	.39	.34	
.075	.65	.71*	.73*	.71	.64	.53	
.100	.91*	.92*	.91*	.86	.78	.63	
			Signed rank test, n = 40				
.000	1.00	1.00	.97	.76	.28	.05	
.025	.99	.92	.59	.14	.07	.42*	
.050	.81	.38	.06	.15	.59*	.93*	
.075	.19	.04	.24	.71*	.96*	1.00*	
.100	.06	.37	.81	.97*	1.00*	1.00*	
			Sign test, n = 40				
.000	1.00*	1.00*	.99*	.84*	.31*	.04	
.025	1.00*	.99*	.82*	.28*	.04	.30	
.050	.98*	.76*	.23	.04	.28	.81	
.075	.67*	.18	.03	.24	.75	.98	
.100	.13	.02	.19	.69	.96	1.00	

*Most powerful of the three tests for that α, β, and n.

TABLE 2

Time of onset of stimulation in hours after hatching and preference at 48 hours after hatching for mallard ducklings

Duckling	Time of onset	Preference*	Duckling	Time of onset	Preference	Duckling	Time of onset	Preference
1	-48.5	N	21	-20.0	S	41	-12.0	S
2	-48.0	N	22	-20.0	S	42	-12.0	S
3	-45.0	N	23	-20.0	N	43	-11.0	S
4	-43.5	N	24	-19.5	N	44	-6.0	N
5	-42.0	S	25	-19.5	S	45	-6.0	N
6	-39.0	S	26	-18.0	N	46	-6.0	N
7	-38.5	S	27	-18.0	N	47	-5.0	S
8	-37.0	N	28	-17.0	S	48	-5.0	S
9	-35.0	S	29	-16.0	N	49	-5.0	S
10	-32.0	N	30	-15.5	N	50	-5.0	S
11	-32.0	N	31	-15.5	N	51	-2.0	S
12	-28.0	S	32	-15.0	N	52	-2.0	N
13	-28.0	N	33	-15.0	N	53	-2.0	S
14	-27.5	N	34	-15.0	N	54	-2.0	N
15	-27.0	N	35	-14.5	N	55	-1.0	N
16	-25.0	S	36	-13.0	S	56	1.0	S
17	-23.5	N	37	-13.0	N	57	2.0	S
18	-22.0	N	38	-13.0	S	58	2.0	N
19	-21.0	N	39	-12.0	N	59	5.0	N
20	-20.5	N	40	-12.0	N	60	6.0	N

*N is normal maternal call; S is slowed maternal call.