Testing change-point in logistic models with covariate measurement error

Yanyuan Ma

Department of Statistics, Texas A&M University

College Station, TX 77845, USA

Abstract

We test the presence of a change of slope in a logistic regression model with covariate measured with errors. Under the null hypothesis of no change-point, estimation of a single intercept and slope can be carried out straightforwardly by various conditional score based methods. If the alternative hypothesis holds and indeed there exists a change-point, estimation becomes more challenging, nevertheless it can still be carried through via semiparametric procedures. However, this does not warrantee a score type of testing procedure due to a degeneration of the estimating equation for the change-point location under the null. The usual Wald type tests fail as well due to another degeneration caused by the singularity of the information matrix. We propose a Wald type test without requiring to estimate the change-point location. Numerical results show the satisfying performance of the proposed testing procedure in terms of both level precision and power.

Key words and phrases: Change-point; Errors-in-variables; Logistic regression; Semiparametric method; Wald test.

Short title: Change-point test

1 Introduction

Consider the model

$$\Pr(Y = 1|X) = h\{\beta_1 + \beta_2 X + \beta_3 (X - c) I(X > c)\}, \tag{1}$$

where $h(t) = 1/(1 + e^{-t})$. When $\beta_3 = 0$, this is a familiar logistic regression model. When $\beta_3 \neq 0$, the linear function has a change of slope at the point c, hence (1) is referred to as a change-point model. Change-point models often arise in various practical situations. For example, in assessing the possibility of a patient cure under a treatment depending on an individual's biomarker, one might suspect that for patients with the biomarker value above a certain threshold, the treatment works more or less effectively. Hence it is often important to verify if this is indeed the case and if so, what is the threshold value. This has become increasingly important in the personalized medicine era since often two or more drugs exist for a same symptom and a more efficient administration strategy is to use a specific medication in the most responsive patient group. Similar examples come from toxicology. It is known that exposure to high does of radon has a direct link with the occurrence of lung cancer. A legitimate question to ask is what exposure level is the critical point where risk increases dramatically, since a complete elimination of the exposure is very unlikely.

Both problems can be modeled as the logistic change-point model given in (1), where as soon as the covariate reaches the change-point c, the rate of the response-covariate dependence, represented by the slope, changes provided that $\beta_3 \neq 0$. Another common aspect of these two examples is that both the biomarker and the radon exposure level can only be measured imprecisely. To improve the precision, often repeated measurements are taken and the average of these measurements are taken as a better measurements. This naturally leads to an errors-in-variables problem, where we assume the measurements are related to the true yet unobservable variable X through

$$W = X + U, \tag{2}$$

and we typically assume U to be normally distributed with mean zero. Taking advantage of additional information such as repeated measurements, other instrumental variables or validation data information, we can estimate the variance of U in the data preprocessing stage. Because of these common practices, from here on, as far as the change-point problem is concerned, we can treat the distribution of U to be known.

The display (1) together with (2) form the central model of our problem. We denote the observations (W_i, Y_i) for $i = 1, \ldots, n$. Our goal is to, firstly, identify if a change-point really exists, and secondly, to identify the location of the change-point if such a change-point indeed is present. We first note that to establish or eliminate the existence of the change-point is a problem of testing $H_0 : \beta_3 = 0$ versus $H_1 : \beta_3 \neq 0$. If $\beta_3 = 0$, (1) simplifies to $\Pr(Y = 1|X) = h(\beta_1 + \beta_2 X)$, which is a familiar linear logistic regression model, and various methods exist to estimate β_1, β_2 without imposing a distributional assumption on X, see Carroll et al. (2006). When $\beta_3 \neq 0$, (1) deviates away from the generalized linear model framework and a sufficient complete statistic no longer exists. This implies that the traditional conditional score type of methods will no longer applicable and we will have to rely on the semiparametric methods developed in Tsiatis and Ma (2004). An additional complexity here is that the location of the change, c is also unknown and the regression function is not smooth with respect to c. Thus, even the semiparametric methods cannot be straightforwardly used without modification.

Even if we are able to perform the estimation both under the null and under the alternative, it is not immediately clear how the testing procedure proceeds. This is caused by a rather special degeneration phenomenon specific to the change-point problem in (1). For example, a natural testing strategy for measurement error models is a score type procedure proposed in Ma et al. (2010). This procedure requires estimating "all" the free parameters under the null model and then plugging them into the estimating equations under the alternative. Obviously, under the null model, the change-point degenerates to a simple linear logistic model, hence the free parameter c is not estimable. This immediately

excludes the score type of testing procedure. A second natural procedure is the Wald type test. In fact, as long as we can estimate the parameters at a root n rate with the usual asymptotic normality property, a Wald test statistic can be easily formed by normalizing the square of the corresponding estimate with its asymptotic variance, and comparing it with the chi-square quantiles. However, a closer look at the structure of the model reveals that this is again not applicable due to the degenerated asymptotic variance-covariance matrix. Specifically, when the parameter β_3 is close to zero, the corresponding estimating equation with respect to c is almost identically zero. This results in a rank deficient derivative matrix, which causes the information matric to become singular. This in turn renders the estimation variance-covariance matrix to be unobtainable, hence the standard Wald test also breaks down.

Based on sound mathematical and statistical intuition, we propose a surprisingly simple treatment to the problem. We explain the main idea and describe the working procedure for both the testing and the estimation of the change-point in Section 2. A simulation study is conducted to illustrate level precision and power of the proposed test procedure and the performance of estimation in Section 3. Finally, in Section 4, we provide some further discussion on the related issues.

2 Methodology

2.1 Estimation

We first provide an overview of the semiparametric estimation procedure for a general regression model with measurement error. We write the regression model as $p_{Y|X}(y|x;\beta)$ where β contains all the unknown parameters. Thus, under null, $\beta = (\beta_1, \beta_2)^T$, and under alternative, $\beta = (\beta_1, \beta_2, \beta_3, c)^T$. Under the assumption that $p_{Y|X}(y|x;\beta)$ is differentiable with respect to β. We perform the following procedure. First, we denote the score function $\partial \log p_{Y|X}(y|x;\beta)/\partial\beta$ evaluated at (X_i, Y_i) as $S^F_\beta(X_i, Y_i;\beta)$. We now adopt a model as the

probability density function (pdf) for X, f_X^*, and use E^* to denote expectations calculated using f_X^*. Note that f_X^* does not need to be the true pdf of X. We then solve for $\alpha(X)$, which is a function that satisfies

$$E[E^*\{\alpha(X)|W,Y\}|X] = E[E^*\{S_\beta^F(X_i,Y_i;\beta)|W,Y\}|X].$$

Finally, we form $\psi(w_i,y_i;\beta)$ through

$$\psi(w_i,y_i;\beta) = E^*\{S_\beta^F(X_i,Y_i;\beta)|W=w_i,Y=y_i\} - E^*\{\alpha(X)|W=w_i,Y=y_i\}$$

for $i=1,\ldots,n$. The estimation of β is subsequently obtained via solving the estimating equation

$$\sum_{i=1}^{n}\psi(w_i,y_i;\beta) = 0.$$

Typically, no closed form of $E^*\{S_\beta^F(X_i,Y_i;\beta)|W,Y\}$ exists even under normal measurement error, therefore, the calculation of the score function and its conditional expectation needs to be carried out numerically. As a consequence, $\alpha(X)$ also has to be obtained numerically. In Tsiatis and Ma (2004), it is shown that the above procedure is guaranteed to yield a consistent estimator for β.

Our difficulty here lies in the fact that our pdf is not differentiable with respect to β. Specifically, the change-point model is non-differentiable with respect to the change-point location c. There are two approaches to circumvent this problem. The first is to replace the sharp change with a smooth change. For example, we consider the model

$$\begin{aligned}&\Pr(Y=1|X)\\ &= h\{\beta_1+\beta_2X+\beta_3(X-c+u)^2/(4u)I(c-u\le X\le c+u)+\beta_3(X-c)I(X>c+u)\},\end{aligned}$$

where u is a small positive constant. It is easy to verify that for any positive u, the above model is continuous and differentiable with respect to c, and when $u\to 0$, the model approaches our original model (1). Therefore we can work with the smooth model with

a small u value instead of the original model. Because the approximation of the smooth model to the original model is a numerical error issue instead of a statistical issue, we can always choose u sufficiently small so that the numerical error is ignorable in comparison to the statistical error.

A second approach is even simpler. Note that the non-differentiability only occurs at $c = X$ and is caused by the difference of the left derivative and right derivative with respect to c. This implies that at all other c values, the score function with respect to c does exist. In fact, when $X \neq c$, the score function with respect to c can be easily written as $S_c(X, Y; \beta) = [Y - h\{\beta_1 + \beta_2 X + \beta_3 (X - c) I(X > c)\}](-\beta_2 I(X > c)$. In the procedure of the estimation, the score function only appears in its integrated form via $E^*\{S_\beta^F(X_i, Y_i; \beta)|W, Y\}$, where the expectation is calculated under a model we mandate. Thus, we can easily propose a model f_X^*, for example, f_X^* is continuous, so that $\{X = c\}$ is a zero measure set hence its value has no impact in the subsequent calculation.

The above two procedures are equally effective in practice and can be used to carry out the estimation of β under the alternative.

2.2 Testing

Our next task is to find a testing procedure that does not suffer from the pitfalls mentioned before. A quick fix of the score type test seems not obvious. In an attempt to fix the Wald test, recall that the breakdown of the Wald test is caused by the singularity of the asymptotic information matrix. To this end, a natural thinking is to give up using the asymptotic property, and use an alternative bootstrap based method to obtain the estimation variance. However, our practical experience indicates that this does not solve the problem either, since the bootstrap estimation of the variance will also be close to be singular. This in turn creates numerical instability in the variance estimation. Since the Wald test statistic is formed via $T = \widehat{\beta}_3^2 / \widehat{\text{var}}(\widehat{\beta}_3)$, a crude estimate of $\text{var}(\widehat{\beta}_3)$ certainly will cause the T to be very different from $\widehat{\beta}_3^2 / \text{var}(\widehat{\beta}_3)$, hence leads to an imprecise testing level.

To overcome the above difficulties, a very simple observation turns out to be beneficial. We point out the fact that if indeed, no change-point exists, then we can fix c at an arbitrary value c_a, model (1) with this fixed c_a value will still have $\beta_3 = 0$. If on the other hand, there is indeed a change-point existing at c, then even if we fix c at an arbitrary value c_a that may be different from c, model (1) with this fixed c_a value will still have $\beta_3 \neq 0$, although the slope change will become smaller no matter $c_a < c$ or $c_a > c$. This motivates us to propose a simple testing procedure. We first arbitrarily fix a value for c, say $c = c_a$. Then, treating c as known, we can perform a test on $\beta_3 = 0$ using, say, a Wald test based on the semiparametric estimation procedure. One can easily form the test statistic $T = \widehat{\beta}_3^2 / \widehat{\text{var}}(\widehat{\beta}_3)$ and compare it against the chi-square distribution with 1 degree of freedom. Formally, the p-value equals $1 - \chi_1^2(T)$, where $\chi_1^2(\cdot)$ represents the cumulative distribution function of the chi-square distribution with one degree of freedom. This test should yield the correct level of significance, while it could suffer some power lost. Intuitively, the further away c_a is from c, the more the loss of power will occur.

If H_0 is not rejected, the procedure is formally completed. We could follow with an estimation of β_1, β_2 is desired. However, if H_0 is rejected, a natural subsequent question is to identify the location of the change-point. This is to estimate c and make inference. It can be performed using the modified semiparametric estimator described in Section 2.1. The estimation and inference is now no longer a pathologic problem in theory because the alternative is established. However, in practice, it can still be difficult depending on how large the value β_3 is. Intuitively, if β_3 is large, the change is more dramatic, hence it is relatively easy to identify where this change happens. On the contrary, if the change is fairly small, then with the additional measurement error, it is very difficult to identify where this slight change happens.

We emphasize here that if the proposed model $f_X^*(x)$ is true, the above procedure yields the optimal estimator for c, hence no improvement is possible. This is equivalent to say that even if unsatisfactory performance of the estimation of c happens, it is an inherent property

of the problem and no method exists that will improve it. On the other hand, if $f_X^*(x)$ is not true, the estimation is still consistent. Obviously, estimating $f_X(x)$ is no easy task. In addition, from our experience, the semiparametric estimation property of the parameter β is often very insensitive to the proposed model $f_X^*(x)$. Thus, in practice, we recommend to simply use the estimation and inference result from the modified semiparametric estimation procedure.

3 Simulation

We conduct a simulation study to illustrate the proposed methodology. In this simulation, we generated the X_i's from a uniform distribution between 0 and 3, and the measurement error U_i's from a normal distribution with mean zero and standard deviation 0.04. We constructed both the null model data set with $\beta = (-1.5, 1)$ and three alternative model data sets. The change-point in all three cases are $c = 1.5$ while the β_3 values are respectively $1.5, 1$ and 0.5, representing a dramatic, moderate and small change of the slope at c. We experimented with a pre-decided c_a value to be $0.75, 1.5$ and 2.25 respectively. A sample size of $n = 1000$ is used and all results are based on 1000 simulations.

The result of the testing and estimation results are illustrated in Tables 1 to 4. From the results under H_0 in Table 1, we can see clearly that the level precision is very close to the nominal whether or not we set $c_a = c$ and whether or not we set f_X^* to the true f_X. This indicates the consistency of the test hence the validity of our method. The results in Table 2 to 4 convey several messages on the power of the test. First. Whether or not we have set $c_a = c$, the test has certain power to detect the alternative. Not surprisingly, when $c_a = c$, the power is largest, and it decays when c_a deviates away from c. Second, when the slope change represented by β_3 is large, the performance is better. This is reflected in the power of the test, in the precision of the estimation variance assessment and in the 95% confidence interval coverage probability. Finally, change-point location is relatively

difficult to estimate (the last element in β is the change-point location c). Even under the alternative, where the singularity of the information matrix is not an issue, the estimation is very unstable. This is especially clear when the slope change is small as in Table 4 or when we target at assessing the tail of the variance estimation as in the 95% coverage results in Tables 2 to 4.

4 Discussion

We have proposed a relatively simple procedure to test the existence of a change-point in the logistic regression context with covariate measurement error. The unique aspect of the test is to require pre-fixing a candidate change-point location and then proceed with this candidate location as if it is the true location.

Although we can establish the consistency of the test, this certainly raises the concern about the power loss. At a first look, this power loss seems avoidable via a usual multiple testiing procedure. For example, we set the change-point at several candidate change-point locations and define a new test statistic to be the maximum of the collection of the test statistics at each of these locations. However, the serious issue here is the null distribution of this new test statistic. The asymptotic distribution of the resulting test statistic is no longer obvious, and a traditional bootstrap method fails to produce a sample distribution either. This is because under covariate measurement error, we are not able to generate data under a null model without observing the X_i's or knowing the distribution of X_i's. One could argue that it is still possible to estimate f_X through a deconvolution procedure, but it is known to yield very slow rate (Carroll and Hall, 1988), hence it is not clear it will actually lead to a gain in the final testing procedure.

A much simpler way exists to reduce the power loss. Specifically, we can choose several candidate changepoints, consider the concatination of the β_3 at these changepoints as a new parameter vector and test the vector value equal to zero. Because the essential idea for

the vector testing is identical to what is presented in the main text while computationally more complex, hence we do not further expand on the details.

Acknowledgment

This work was supported by the National Science Foundation (DMS-0906341) and the National Institute of Neurological Disorders and Stroke (R01-NS073671).

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: The Johns Hopkins University Press.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvoluting a density. *Journal of the American Statistical Association*, **83**, 1184-1186.

Carroll, R. J., Ruppert, A., Stefanski, L. A. and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition. London: CRC Press.

Ma, Y., Hart, J. D., Janicki, R. and Carroll, R. J. (2010) Local and Omnibus Tests in Classical Measurement Error Models. *Journal of the Royal Statistical Society, Series B*, in press.

Tsiatis, A. A. and Ma, Y. (2004) Locally Efficient Semiparametric Estimators for Functional Measurement Error Models. *Biometrika*, **91**, 835-848.

Table 1: Level precision of the test and estimation results in the simulation. True parameter values (β_0), average of the estimated parameter values ($\widehat{\beta}$), empirical standard deviation (sd), average of the estimated standard errors ($\widehat{\text{sd}}$) and coverage probability of 95% confidence intervals (95%) are reported. Sample size $n = 1000$, results based on 1000 simulations.

	Uniform f_X^*				Normal f_X^*			
	Testing							
c_a	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2
0.75	0.010	0.052	0.104	0.206	0.010	0.051	0.107	0.205
1.5	0.008	0.051	0.108	0.207	0.006	0.052	0.108	0.208
2.25	0.012	0.050	0.098	0.209	0.013	0.051	0.100	0.208
	Estimation							
β_0	-1.5	1	-	-	-1.5	1	-	-
$\widehat{\beta}$	-1.5061	1.0017	-	-	-1.5087	1.0033	-	-
sd	0.1469	0.0877	-	-	0.1472	0.0878	-	-
$\widehat{\text{sd}}$	0.1456	0.0855	-	-	0.1458	0.0857	-	-
95%	95.1%	95.3%	-	-	94.9%	95.3%	-	-

Table 2: Power of the test and estimation results in the simulation for large slope change. True parameter values (β_0), average of the estimated parameter values ($\widehat{\beta}$), empirical standard deviation (sd), average of the estimated standard errors ($\widehat{\text{sd}}$) and coverage probability of 95% confidence intervals (95%) are reported. Sample size $n = 1000$, results based on 1000 simulations.

	Uniform f_X^*				Normal f_X^*			
	Testing							
c_a	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2
0.75	0.592	0.785	0.859	0.920	0.588	0.778	0.856	0.921
1.5	0.869	0.953	0.979	0.992	0.867	0.955	0.979	0.992
2.25	0.319	0.677	0.802	0.903	0.313	0.673	0.801	0.899
	Estimation							
β_0	-1.5	1	1.5	1.5	-1.5	1	1.5	1.5
$\widehat{\beta}$	-1.4967	0.9824	1.6089	1.5193	-1.4985	0.9832	1.6205	1.5207
sd	0.2096	0.2313	0.4494	0.1739	0.2105	0.2339	0.5033	0.1824
$\widehat{\text{sd}}$	0.2080	0.2246	0.4574	0.1796	0.2084	0.2250	0.5326	0.1737
95%	94.3%	94.6%	96.5%	59.4%	93.7%	94.1%	96.7%	56.7%

Table 3: Power of the test and estimation results in the simulation for moderate slope change. True parameter values (β_0), average of the estimated parameter values ($\widehat{\beta}$), empirical standard deviation (sd), average of the estimated standard errors ($\widehat{\text{sd}}$) and coverage probability of 95% confidence intervals (95%) are reported. Sample size $n = 1000$, results based on 1000 simulations.

	Uniform f_X^*				Normal f_X^*			
	Testing							
c_a	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2
0.75	0.289	0.528	0.651	0.770	0.284	0.525	0.646	0.767
1.5	0.558	0.779	0.873	0.927	0.560	0.781	0.871	0.927
2.25	0.167	0.466	0.621	0.759	0.167	0.457	0.617	0.754
	Estimation							
β_0	-1.5	1	1	1.5	-1.5	1	1	1.5
$\widehat{\beta}$	-1.4927	0.9760	1.0988	1.5240	-1.4961	0.9781	1.0993	1.5214
sd	0.2034	0.2310	0.3955	0.2195	0.2047	0.2233	0.3980	0.2184
$\widehat{\text{sd}}$	0.2070	0.2262	0.4177	0.2610	0.2077	0.2260	0.4203	0.2529
95%	94.5%	94.4%	96.5%	54.6%	94.8%	94.3%	96.6%	54.7%

Table 4: Power of the test and estimation results in the simulation for small slope change. True parameter values (β_0), average of the estimated parameter values ($\widehat{\beta}$), empirical standard deviation (sd), average of the estimated standard errors ($\widehat{\text{sd}}$) and coverage probability of 95% confidence intervals (95%) are reported. Sample size $n = 1000$, results based on 1000 simulations.

	Uniform f_X^*				Normal f_X^*			
	Testing							
c_a	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2
0.75	0.066	0.190	0.277	0.423	0.064	0.189	0.275	0.419
1.5	0.127	0.310	0.424	0.586	0.123	0.307	0.423	0.585
2.25	0.050	0.184	0.281	0.429	0.051	0.180	0.277	0.426
	Estimation							
β_0	-1.5	1	0.5	1.5	-1.5	1	0.5	1.5
$\widehat{\beta}$	-1.5005	0.9884	0.5545	1.5266	-1.5044	0.9918	0.5446	1.5225
sd	0.2007	0.2093	0.3882	0.2905	0.2005	0.2080	0.3854	0.2703
$\widehat{\text{sd}}$	0.2061	0.2274	0.4107	0.7578	0.2090	0.2213	0.5056	0.8558
95%	93.9%	92.0%	91.4%	44.0%	94.3%	92.0%	91.0%	44.7%