Journal of the Korean
Statistical Society
Vol. 21, No. 1, 1992

BOOTSTRAPPING GENERALIZED LINEAR MODELS WITH RANDOM REGRESSORS [1]

Kee-Won Lee, [2] Choongrak Kim, [3] Keon Tae Sohn[4] and Kwang Mo Jeong[5]

ABSTRACT

The generalized linear models with random regressors case are studied for bootstrapping. Only the natural link functions are considered. It is shown that the bootstrap approximation to the distribution of the maximum likelihood estimators is valid for almost all sample sequences. A slight extension of this model is also considered.

1. INTRODUCTION

Some asymptotic theory for applications of Efron's(1979) bootstrap to generalized linear models with random regressors case is considered. For an extensive review of generalized linear models, see, for example, McCullagh and Nelder(1989). For consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, see Fahrmeir and Kaufmann(1985).

The bootstrap approximation to the distribution of the least squares estimates in linear model context is studied by Freedman(1981), and the less metrical version

[1] This research is supported from Korea Ministry of Education(BSRI-91-116).

[2] Department of Statistics, Hallym University, Chunchon, Kangwon, 200-702, Korea.
[3] Department of Statistics, Pusan University, Pusan, 609-735, Korea.
[4] Department of Statistics, Pusan University, Pusan, 609-735, Korea.
[5] Department of Statistics, Pusan University, Pusan, 609-735, Korea.

of the same result can be found in Beran(1984), which gives more examples. Nice reviews of bootstrap methods developed so far are given by Hinkley(1988), and DiCiccio and Romano(1988), where many references are cited. In this paper some asymptotic results will be given in the context of generalized linear models with random regressors case. Our attention will be restricted to the case of natural link functions.

Section 2 introduces the model with some asymptotic properties of the estimators, and summarizes the way of bootstrapping and its validity. Section 3 introduces a slight extension of this model and some asymptotic properties of the estimators, and the way of bootstrapping and its validity. Even though the estimators are algebraically the same, different probability model leads to the different asymptotic results. Proofs are sketched in Section 4.

2. MODEL, MLE, AND THE BOOTSTRAP

Suppose we have n independent and identically distributed random vectors (Y_i, X_i^t) for $i = 1, 2, \ldots, n$ such that the covariates X's are p-variate random vectors with distribution function $G(x)$, which is unknown, and there exists a positive number M such that $\|X\| \leq M$ with probability 1.

Given $X = x$, the response Y belongs to a natural exponential family with density given by

$$f(y|x) = c(y)\exp[y\theta - b(\theta)], \tag{2.1}$$

where θ belongs to a natural parameter space Θ in R^1. For more properties of natural exponential families, see Lehmann(1986).

The random regressor X is connected to the response Y through a natural link function $\theta = X^t\beta$, where β is a p-vector of unknown parameters. Our attention will be restricted to β's such that $X^t\beta$ belongs to Θ with probability 1. Throughout the paper *w.p.*1. will be an abbreviation for *with probability* 1, which refers to the random mechanism generating the original sample. In the sequel, β_0 denotes the true but unknown parameter, which is supposed to generate our observations.

The log-likelihood for a single observation is given by

$$\ell(\beta,\ G;\ y,\ x) = yx^t\beta - b(x^t\beta) + \text{constants}. \tag{2.2}$$

It can be shown, by chain rule, that the the first derivative of ℓ with respect to β is given by

$$\nabla \ell = x(y-\mu), \tag{2.3}$$

where $\mu = \nabla b(x^t\beta) = E(Y|X=x)$. Hence the first derivative does not depend on the unknown cumulative distribution function G, so the efficient score function for estimating β when G is unknown is the same as that when G is known. Therefore the solution of the likelihood equation

$$\sum x_i(y_i-\mu_i)=0, \quad \mu_i = \nabla b(x_i^t\beta), \tag{2.4}$$

which can be obtained by any iterative method is asymptotically efficient. McCullagh and Nelder(1989) describes one typical algorithm in detail. Here $\sum$ denotes the summation over all the observations. We have the matrix of second derivatives

$$\nabla^2\ \ell = -x\ x^t\sigma^2, \text{ where } \sigma^2 = \nabla^2 b(x^t\beta) = Var(Y|X=x). \tag{2.5}$$

So the information matrix regarding β, denoted by $I(\beta)$, is given by

$$E[\nabla\ell(\nabla\ell)^t] = E[XX^t(Y-\mu)^2]. \tag{2.6}$$

Or, equivalently,

$$-E(\nabla^2\ell) = E(XX^t\sigma^2). \tag{2.7}$$

Now we have the following asymptotic properties for the solution of the likelihood equation, which are easy consequence of more general results given in Fahrmeir and Kaufmann(1985).

Lemma 2.1. In addition to the assumptions regarding the model just described above, suppose that there does not exist a vector of constants $a=(a_1,\ldots,a_p)$ and a real number c such that $aX=c$ w.p.1. Then the solution of the likelihood equation

$$\sum x_i(y_i-\mu_i)=0, \quad \mu_i = \nabla b(x_i^t\beta), \tag{2.8}$$

which is unique if it exists, satisfies the following;

$$\hat\beta_n - \beta_0 \to 0 \quad \text{w.p.1}, \tag{2.9}$$

where $\hat\beta_n$ is the solution of the likelihood equation, furthermore,

$$n^{\frac{1}{2}}(\hat\beta_n - \beta_0) \to N_p(0, I^{-1}(\beta_0)) \text{ in distribution.} \tag{2.10}$$

Remark It can be easily checked that the additional condition is needed to guarantee that the information matrix $I(\beta_0)$ is invertible.

There are at least two ways of estimating $I(\beta_0)$. One uses the first expression for the information matrix, and the other uses the second expression, which gives the same matrix under our model. Let

$$\hat{I} = \frac{\sum X_i X_i^t \sigma_i^2}{n}, \tag{2.11}$$

and let

$$\tilde{I} = \frac{\sum X_i X_i^t (Y_i - \hat{\mu}_i)^2}{n}, \text{ where } \hat{\mu}_i = \nabla b(X_i^t \hat{\beta}_n),\ \hat{\sigma}_i^2 = \nabla^2 b(X_i^t \hat{\beta}_n). \tag{2.12}$$

We have the following lemma whose proof will be sketched in section 4.

Lemma 2.2. Under the assumptions of Lemma 2.1, both estimators of the information matrix $I(\beta_0)$ are strongly consistent. Therefore studentized versions of second part of Lemma 2.1 also hold by Slutsky's theorem.

Now for the bootstrap, note that we have an independent and identically distributed probability structure in this case and the unknown parameters are β_0 and $G(\cdot)$, and they can be consistently estimated by the MLE $\hat{\beta}_n$ of β_0 and the empirical cdf based on X's. The bootstrap goes as follows.

Step 1. From the original sample X's and Y's, compute the MLE of β_0 by any iterative method and construct an empirical cdf based on X's only.

Step 2. Generate bootstrap sample X^*'s and Y^*'s as follows. X^*'s are a simple random sample with replacement from the original X's. Given $X^* = x^*$, choose Y^* from a natural exponential family with density given by

$$f(y|x^*) = c(y) \exp[y\hat{\theta}^* - b(\hat{\theta}^*)], \text{ where } \hat{\theta}^* = x^{*t}\hat{\beta}_n. \tag{2.13}$$

Step 3. From the bootstrap sample, compute $\hat{\beta}_n^*$, which maximizes the bootstrap version of the log-likelihood, say ℓ^*. Explicitly, $\hat{\beta}_n^*$ is the solution to the equation,

$$\nabla \ell^* = \sum X_i^*(Y_i^* - \mu_i^*) = 0, \text{ where } \mu_i^* = \nabla b(X_i^{*t}). \tag{2.14}$$

The following theorem tells us that the above bootstrap approximation is valid for almost all sample sequences.

Theorem 1. Assume that $E(XX^t)$ exists and is positive definite in addition to the assumtions of Lemmma 2.1. Then, given the original sample, for almost all sample sequences, all the bootstrap versions of the previous lemmas hold. That is,

$$n^{\frac{1}{2}}(\hat{\beta}_n^* - \hat{\beta}_n) \longrightarrow N_p(0, I^{-1}(\beta_0)) \text{ in distribution,} \tag{2.15}$$

$$\hat{I}^*(\hat{\beta}_n^*) \longrightarrow I(\beta_0) \text{ in conditional probability,} \tag{2.16}$$

$$\tilde{I}^*(\hat{\beta}_n^*) \longrightarrow I(\beta_0) \text{ in conditional probability,} \tag{2.17}$$

$$n^{\frac{1}{2}}\hat{I}^{*\frac{1}{2}}(\hat{\beta}_n^*)(\hat{\beta}_n^* - \hat{\beta}_n) \longrightarrow N_p(0, I_p) \text{ in distribution,} \tag{2.18}$$

$$n^{\frac{1}{2}}\tilde{I}^{*\frac{1}{2}}(\hat{\beta}_n^*)(\hat{\beta}_n^* - \hat{\beta}_n) \longrightarrow N_p(0, I_p) \text{ in distribution,} \tag{2.19}$$

where I_p denotes the $p \times p$ identity matrix, and the starred items denote the bootstrap versions of the original ones, for example,

$$\hat{I}^*(\beta) = \sum X_i^* X_i^{*t} \nabla^2 b(X_i^{*t}\beta), \text{ etc.} \tag{2.20}$$

3. AN EXTENSION

In this section, a slight extension of the model described in section 2 is considered. Regard all of the n observations, (Y_i, X_i^t) for $i = 1, \ldots, n$ as independent and identically distributed according to a $(p+1)$ dimensional distribution function F such that there exists a positive number M satisfying $\|X\| \leq M$ with probability 1 and the 4th moment of Y exists.

If we use a model selection terminology, this F belongs to an operating family, and the generalized linear model structure discussed in the previous section now becomes an approximating family. Linhart and Zucchini(1986) gives a good introduction to model selection.

Therefore the only unknown parameter in this case is F, and the functional $\beta_0(F)$ is a minimizer such that the Kullback-Leibler divergence, which is a natural choice for measuring discrepancy between the operating family and the approximating family in our case, between F and the previous generalized linear model structure, takes its minimum. That is to say,

$$\beta_0(F) = \arg\min_{\beta}\Big\{ - E_F[\ell(\beta)]\Big\}. \tag{3.1}$$

A natural and consistent estimator of the above discrepancy, called an empirical discrepancy, is the one with an empirical cdf, call it F_n, in place of F. Now that estimator of the discrepancy coincides with minus times the log-likelihood described in the previous section.

An empirical version of $\beta_0(F)$, which gives a minimum of the empirical discrepancy, coincides analytically with the maximum likelihood estimator described in section 2. But, since we have different probability model, different asymptotic results will be developed unless F includes the generalized linear model described in the previous section. For the simplicity of notation, the same notations will be used as in the previous section, but we have to keep it in mind that we are working on a different probability model. Now let

$$\Omega = -E_F(\nabla^2\ell), \quad \sum = E_F[\nabla\ell(\nabla\ell)^t]. \tag{3.2}$$

Note that $\Omega = \sum$ when F coincides with the generalized linear model. Now we have the following lemma which is parallel to the Lemma 2.1. For the regularity conditions and the proof, see Huber(1967), or the appendix of Linhart and Zucchini(1986).

Lemma 3.1. Under proper conditions, which can be checked using the assumptions on F, the following holds.

$$\hat\beta_n \to \beta_0 \quad \text{w.p.1}, \tag{3.3}$$

$$n^{\frac{1}{2}}(\hat\beta_n - \beta_0) \to N_p(0, \Omega^{-1}\Sigma\Omega^{-1}) \text{ in distribution}, \tag{3.4}$$

where β_0 is an abbreviation for $\beta_0(F)$, and $\hat\beta_n = \beta_0(F_n)$.

Natural estimators of Ω and Σ are

$$\Omega_n(\hat\beta_n) = \frac{\sum X_i X_i^t \hat\sigma_i^2}{n}, \tag{3.5}$$

and

$$\Sigma_n(\hat\beta_n) = \frac{\sum X_i X_i^t (Y_i - \hat\mu_i)^2}{n}, \tag{3.6}$$

where $\hat\mu_i = \nabla b(X_i^t\hat\beta_n)$, $\hat\sigma_i^2 = \nabla^2 b(X_i^t\hat\beta_n)$, respectively. Both of these were used to estimate the information matrix in section 2.

Lemma 3.2. Both of the above estimators are strongly consistent. Therefore a studentized version of the second part of Lemma 3.1 also holds.

For the bootstrap, note that the only unknown parameter is F, which can be consistently estimated by the empirical cumulative distribution function F_n. The bootstrap goes as follows.

Step 1. From the original sample, compute $\hat{\beta}_n$ and construct an empirical cumulative distribution function F_n based on the whole sample, that is, assign probability mass of $\frac{1}{n}$ to each of the observations, (Y_i, X_i^t) for $i = 1, 2, \dots, n$.

Step 2. Choose a bootstrap sample from the original sample, that is, a simple random sample with replacement from the original sample of the same size. For notational simplicity write them. (Y_i^*, X_i^{*t}) instead of $(Y_i, X_i^t)^*$.

Step 3. Compute a bootstrap version of $\hat{\beta}_n$, say $\hat{\beta}_n^*$ from the bootstrap sample.

Theorem 2. The following holds under the assumptions of the Lemma 3.2. For almost all sample sequences, given the original sample, as n tends to infinity,

$$n^{\frac{1}{2}}(\hat{\beta}_n^* - \hat{\beta}_n) \rightarrow N_p(0, \Omega^{-1}\Sigma\Omega^{-1}) \text{ in distribution,} \tag{3.7}$$

$$\Omega_n^*(\hat{\beta}_n^*) \rightarrow \Omega \text{ in conditional probability,} \tag{3.8}$$

$$\Sigma_n^*(\hat{\beta}_n^*) \rightarrow \Sigma \text{ in conditional probability,} \tag{3.9}$$

$$n^{\frac{1}{2}}\Omega_n^{*\frac{1}{2}}(\hat{\beta}_n^*)\Sigma_n^{*-\frac{1}{2}}(\hat{\beta}_n^*)\Omega_n^{*\frac{1}{2}}(\hat{\beta}_n^*)(\hat{\beta}_n^* - \hat{\beta}_n) \rightarrow N_p(0, I_p) \text{ in distribution,} \tag{3.10}$$

where $\Omega_n^*(\beta) = \dfrac{\sum X_i^* X_i^{*t} \sigma_i^{*2}}{n}$, $\Sigma_n^*(\beta) = \dfrac{\sum X_i^* X_i^{*t}(Y_i^* - \mu_i^*)^2}{n}$, $\mu_i^* = \nabla b(X_i^{*t}\beta)$
and $\sigma_i^{*2} = \nabla^2 b(X_i^{*t}\beta)$.

4. SKETCH OF THE PROOFS

4.1. Proof of Lemma 2.2.

Let $T_n(\beta) = \dfrac{\sum X_i X_i^t \sigma_i^2}{n}$, with $\sigma_i^2 = \nabla^2 b(X_i\beta)$, then we may write

$$\hat{I} - I(\beta_0) = [\hat{I} - T_n(\beta_0)] + [T_n(\beta_0) - I(\beta_0)]. \tag{4.1.1}$$

The second term of the right side hand side tends to a matrix of zeroes by the strong law of large numbers. On the other hand the first term of the right hand side also tends to a matrix of zeroes as follows. The (j,k)th element of the first term can be written as, using Taylor expansion and the analytic property of the function $b(\cdot)$,

$$\begin{aligned} |[\hat{I} - I(\beta_0)]_{j,k}| &= |(\hat{\beta}_n - \beta_0)^t \sum X_{ij} X_{ik} X_i \nabla^3 b(X_i^t \tilde{\beta}_n)/n| \\ &\leq (K_3/3\sqrt{3}) \|\hat{\beta}_n - \beta_0\| \sum \|X_i\|^3/n \\ &\to 0, \quad w.p.1, \end{aligned} \tag{4.1.2}$$

where $\|\tilde{\beta}_n - \beta_0\| \leq \|\hat{\beta}_n - \beta_0\|$ and for some positive number K_3 such that

$$\sup |\nabla^3 b(X_i \beta)| \leq K_3 \quad \text{for all } \beta \text{ with } \|\beta - \beta_0\| \leq \epsilon. \tag{4.1.3}$$

The second part of the lemma can be shown by a similar argument to the one given above. Note also that a similar argument can be used to prove Lemma 3.2.

4.2. Proof of Theorem 1.

Let $\ell^*(\beta)$ be the log-likelihood based on the bootstrap sample, that is,

$$\ell^*(\beta) = \sum y_i^* x_i^{*t} \beta - b(x_i^{*t} \beta). \tag{4.2.1}$$

The following expansion will be used.

$$\nabla \ell^*(\hat{\beta}_n^*) = \nabla \ell^*(\hat{\beta}_n) + (\hat{\beta}_n^* - \hat{\beta}_n)^t \nabla^2 \ell^*(\tilde{\beta}_n^*), \tag{4.2.2}$$

where $\|\tilde{\beta}_n^* - \hat{\beta}_n\| \leq \|\hat{\beta}_n^* - \hat{\beta}_n\|$.

Using the fact that $\nabla \ell^*(\hat{\beta}_n^*) = 0$, the above equation can be rewritten as

$$\nabla \ell^*(\hat{\beta}_n)/n^{\frac{1}{2}} = n^{\frac{1}{2}} (\hat{\beta}_n^* - \hat{\beta}_n)^t [-\nabla^2 \ell^*(\tilde{\beta}_n^*)/n]. \tag{4.2.3}$$

For any vector of constants $a = (a_1, \ldots, a_p)$, we may write

$$a^t \nabla \ell^*(\hat{\beta}_n)/n^{\frac{1}{2}} = \sum U_{in}^*, \tag{4.2.4}$$

where $U_{in}^* = a^t X_i^*(Y_i^* - \hat{\mu}_i^*)/n^{\frac{1}{2}}$ with $\hat{\mu}_i^* = \nabla b(X_i^{*t} \hat{\beta}_n)$.

It can be easily checked that the conditional mean and variace of U_{in}^* given the original samples are

$$E(U_{in}^*) = 0, \; Var(U_{in}^*) = \frac{a^t \hat{I} a}{n} \text{ respectively.} \tag{4.2.5}$$

Furthermore, we can find a positive number C_3 such that the third absolute moment of Y given X is bounded by C_3 in a neighborhood of β_0 by the analytic property of $b(\cdot)$. Then

$$E(|U_{in}^*|^3) \leq C_3/n^{\frac{3}{2}} \|a\|^3 \sum \|X_i\|^3/n. \tag{4.2.6}$$

Therefore

$$E(|U_{in}^*|^3)/s_n^3 \rightarrow 0 \qquad \text{w.p.1}, \tag{4.2.7}$$

where $s_n^2 = \sum Var(U_{in}^*) = a^t \hat{I} a$.

Since $a^t \hat{I}$ a tends to $a^t I(\beta_0)$ a with probability 1, the Lyapounov's condition for a triangular array is checked and we have established that, with probability 1,

$$\nabla \ell^*(\hat{\beta}_n)/n^{\frac{1}{2}} \rightarrow N_p(0, I(\beta_0)) \text{ in distribution.} \tag{4.2.8}$$

By a similar argument, with the aid of weak law of large numbers for triangular array, a bit more complicated though, to the proof of lemma 2.2, it can be shown that, with probability 1,

$$-\nabla^2 \ell^*(\hat{\beta}_n)/n \rightarrow I(\beta_0) \text{ in conditional probability.} \tag{4.2.9}$$

Now (4.2.3), (4.2.8), (4.2.9), and Lemma 6.4.1 in Lehmann(1983) completes the proof of the theorem.

4.3. Proof of Theorem 2.

Proof parallels that of Theorem 1. The same steps can be followed except in (4.2.8) and (4.2.9) in the proof. The difference can be summarized as;

$$\nabla \ell^*(\hat{\beta}_n)/n^{\frac{1}{2}} \rightarrow N_p(0, \Sigma) \text{ in distribution.} \tag{4.2.8}$$

$$-\nabla^2 \ell^*(\hat{\beta}_n)/n \rightarrow \Omega \text{ in conditional probability.} \tag{4.2.9}$$

and we will get the desired result.

REFERENCES

(1) Beran, R.J.(1984). Bootstrap methods in statistics. *Jahresbericht der Deutschen Mathematiker - Vereiningung*, 86, 14 - 30.

(2) DiCiccio,T.J. and Romano, J.P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society*, B 151, 321 - 337.

(3) Efron, B.(1979). Bootstrap methods. *Annals of Statistics*, 7, 1 - 26.

(4) Fahrmeir and Kaufmann, H.(1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics*, 13, 342 - 368.

(5) Freedman, D.A.(1981). Bootstrapping regression models. *Annals of Statistics*, 9, 1218 - 1228.

(6) Hinkley, D.(1988). The bootstrap. *Journal of the Royal Statistical Society*, B 151, 338 - 354.

(7) Huber, P.(1967). The behaviour of maximum likelihood estimates under nonstandard conditions, *Proceedings of the Fifth Berkeley Simposium on Mathematical Statistics and Probability, Vol. 1*, University of California Press, Berkeley.

(8) Huber, P.(1981). *Robust Statistics*, J. Wiley & Sons, New York.

(9) Lehmann, E.L.(1983). *Theory of Point Estimation*, J. Wiley & Sons, New York.

(10) Linhart, H. and Zucchini, W.(1986). *Model Selection*, J. Wiley & Sons, New York.

(11) McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2nd ed., Chapman and Hall, London.