Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–14

DOI: 10.1177/0962280215602716
smm.sagepub.com
SAGE

Natural interpretations in Tobit regression models using marginal estimation methods

Wei Wang and Michael E Griswold

Abstract

The Tobit model, also known as a censored regression model to account for left- and/or right-censoring in the dependent variable, has been used in many areas of applications, including dental health, medical research and economics. The reported Tobit model coefficient allows estimation and inference of an exposure effect on the latent dependent variable. However, this model does not directly provide overall exposure effects estimation on the original outcome scale. We propose a *direct-marginalization* approach using a reparameterized link function to model exposure and covariate effects directly on the truncated dependent variable mean. We also discuss an alternative *average-predicted-value*, post-estimation approach which uses model-predicted values for each person in a designated reference group under different exposure statuses to estimate covariate-adjusted overall exposure effects. Simulation studies were conducted to show the unbiasedness and robustness properties for both approaches under various scenarios. Robustness appears to diminish when covariates with substantial effects are imbalanced between exposure groups; we outline an approach for model choice based on information criterion fit statistics. The methods are applied to the Genetic Epidemiology Network of Arteriopathy (GENOA) cohort study to assess associations between obesity and cognitive function in the non-Hispanic white participants.

Keywords

Tobit model, censored, *direct-marginalization* approach, *average-predicted-value* approach, overall exposure effects

1 Introduction

Often in randomized clinical trials and observational cohort studies, the outcome of interest may be limited due to various inherent censoring mechanisms (i.e. an output with a specified range or a measurement with an inherent detection limit). For example, in the Jackson Heart Study to investigate the causes of cardiovascular disease (CVD) in African Americans, Fox et al. examined the relation of plasma B-type natriuretic peptide (BNP) to body mass index in 3742 participants, and

Center of Biostatistics and Bioinformatics, University of Mississippi Medical Center, Jackson, MS, USA

Corresponding author:
Wei Wang, Center of Biostatistics and Bioinformatics, New Guyton Research Building G562, University of Mississippi Medical Center, 2500 North State Street, Jackson, MS 39216, USA.
Email: wwang@umc.edu

the minimal detectable concentration of BNP with a chemiluminescent immunoassay was 2.0 pg/mL.[1] A major problem in the analysis of such response variables is that the outcome does not attain any of the common distributions available in standard generalized linear modeling approaches. Ignoring boundary issues can cause both bias and incorrect standard error estimates.[2] In cross-sectional studies (especially in econometrics), a common solution applied to this problem is the so-called Tobit model, after Tobin's classical example on household expenditures with a high preponderance of zero-values.[3] The Tobit model, a special case of the more general censored regression model which uses a truncated normal distribution with inflation at the censoring point, has been used in many areas of biomedical science; for example, cytokine concentrations,[4] viral loads subject to detection limits,[5,6] or cognitive function measurements with pre-defined boundaries.[7,8] A motivating example is provided by the GENOA Cohort Study,[9] where associations between obesity and cognitive dysfunction are examined using the 30-point analog scale Mini-Mental State Examination (MMSE); an overly large proportion of subjects in this study recorded upper-limit scores of 30. The problem presented thus involves the assessment of the overall obesity effect on the MMSE score using the Tobit regression model.

Much of the Tobit model literature has focused on models incorporating the censoring nature of the data, including extended Tobit models involving random effects for clustered and longitudinal data,[10,11] weighted random effects Tobit regression models to account for simultaneous non-ignorable missingness and left-censoring,[12] and Bayesian semi-parametric Tobit mixture models.[13] Typically, papers presenting results from a Tobit regression analysis show estimation and inference of an exposure effect on the latent dependent variables[1,14]; however, the latent exposure effect may differ substantially from the observed marginal effect, leading to potentially inappropriate translations of the exposure effect magnitude. So, it would often be more appropriate, interpretable, and valuable to show exposure effects based on the original (versus latent) response variable. One proposed approach for estimating marginal effects from Tobit models involves disaggregating the total changes in the censored outcome into changes in the probability of attaining noncensored values and changes in the noncensored values themselves,[15] but the usage of this approach is limited when the Tobit model includes baseline covariates[16] and alternative approaches are needed.

This article presents two methods for assessing exposure effects on the overall mean in the context of Tobit regression models involving different baseline covariates. The first method estimates these effects using a *direct-marginalization* approach in which estimates have been marginalized to the overall mean by reparameterizing the link function. This approach can be applied by marginalizing over the truncated normal space and boundary components of the response variable to obtain the overall effects estimate. Long et al. proposed a similar method in the context of zero-inflated Poisson (ZIP) models.[17] Estimation the ZIP context is straightforward, since the marginal mean is a simple linear function of the latent mean and thus does not require sophisticated numerical solving algorithms. In the Tobit context we describe here, the relationship between the latent and marginal means is a complex nonlinear function, and advanced root-finding methods such as Newton-Raphson[18] or Brent's methods[19] are needed in order to utilize the original Tobit likelihood. Once solved though, advantages of full likelihood methods automatically translate to our approach, including profile-likelihood options, direct Bayesian extensions by specifying prior distributions for parameters and providing valid inferences when data are missing at random (MAR).[20,21] This technique has also been applied in marginalized random effects models (MREM) for longitudinal categorical data[22] and zero-inflated clustered count data.[20] The Fisher-information matrix is used to calculate the covariance matrix associated with maximum-likelihood estimates.[23]

We have additionally developed an alternative approach to estimate marginal effects denoted as an '*average-predicted-value*' (*APV*) method, which uses model-predicted values for each person under different exposure statuses to assess an overall mean exposure effect in the context of zero-inflated regression models.[24] This idea has been used by Greenland[25] and Localio AR et al.[26] for estimation of a relative risk, Bender et al.[27] to compare the number needed to treat between groups based on a logistic regression model, and Austin[28] for estimating the odds difference for a binary outcome assuming a logistic regression model. Our *APV* approach may be flexibly applied to contrast any function of the overall response means between exposure groups. Whenever expected values under different exposure statuses can be expressed as a function of estimable parameters, the *APV* method can be implemented for Tobit regression models by using the corresponding expected mean expression form to compare marginal response means between different exposure groups. For this *APV* approach, we actually have two levels of marginalization, the first level is to marginalize the latent components of the Tobit model to get a marginal mean for each individual which is conditional on their covariate profile, and the second level is to marginalize over the covariate distribution of subjects in the reference population to compare different exposure groups. Delta methods can be used for variance estimation in this approach.

The rest of the paper is organized as follows. Section 2 describes the Tobit regression model. In Section 3, we detail our *direct-marginalization* and *APV* approaches for inference on overall exposure effects in the Tobit regression model. Section 4 presents a simulation study that compares the alternative methods in terms of bias, efficiency and coverage of the confidence intervals. In Section 5, we apply the new methods to assess effects of obesity on MMSE cognitive scores in GENOA Cohort Study Non-Hispanic Whites. Section 6 provides discussion and concluding remarks.

2 Statistical models

We consider models for a continuous outcome y with upper and/or lower limits based on the Tobit regression model. The general idea of Tobit regression is that it models both the probability of reaching either the lower or upper limit and the probability in-between. Let y^* be a latent random variable that is not censored. The standard Tobit model specifies a linear regression on the latent scale

$$y_i^* = \boldsymbol{X_i'}\boldsymbol{\beta}^* + \varepsilon_i, \quad \varepsilon_i \sim N(0,\ \sigma^2) \tag{1}$$

where i refers to subject i and $\boldsymbol{\beta}^*$ is a fixed effect regression coefficient vector for the covariate vector $\boldsymbol{X_i}$, where $\boldsymbol{X_i}$ includes a primary treatment\exposure T and subsets of other covariates $\boldsymbol{W}$. Furthermore, it is assumed that we can observe the realizations of y^* for a given range $[l, u]$ only, and that values of y^* smaller than l or larger than u are censored at, respectively, l and u. Hence, the observed limited dependent variable y is obtained from y^* as

$$y_i = \begin{cases} l & \text{for } y_i^* \le l \\ y_i^* & \text{for } l < y_i^* < u \\ u & \text{for } y_i^* \ge u \end{cases} \tag{2}$$

If a dependent variable is limited on a single side, only a lower (or upper) limit is needed ($l=-\infty$ or $u=+\infty$). The likelihood for this model is

$$L=\prod_{y_i=l}\left[\Phi\left(\frac{l-\mu_i^*}{\sigma}\right)\right]\prod_{l<y_i<u}\left[\frac{1}{\sigma}\phi\left(\frac{y_i-\mu_i^*}{\sigma}\right)\right]\prod_{y_i=u}\left[1-\Phi\left(\frac{u-\mu_i^*}{\sigma}\right)\right] \tag{3}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal probability and cumulative density functions, and $E(y_i^*)=\mu_i^*=\boldsymbol{X_i'\beta^*}$. Due to the censoring mechanism, the distribution for the observed outcome on its natural scale, y_i, is not the same as the distribution for the assumed latent variable y_i^* and $E(y_i)=\mu_i$ is not equal to $E(y_i^*)=\mu_i^*$. However, the natural scale mean can be expressed as

$$\begin{aligned}\mu_i=\Delta(\mu_i^*)&=\left[\Phi\left(\frac{u-\mu_i^*}{\sigma}\right)-\Phi\left(\frac{l-\mu_i^*}{\sigma}\right)\right]\\&\times\mu_i^*-\left[\phi\left(\frac{u-\mu_i^*}{\sigma}\right)-\phi\left(\frac{l-\mu_i^*}{\sigma}\right)\right]\times\sigma\\&+\Phi\left(\frac{l-\mu_i^*}{\sigma}\right)\times l+\Phi\left(\frac{\mu_i^*-u}{\sigma}\right)\times u\end{aligned} \tag{4}$$

The log-likelihood of the Tobit model is defined as the log of the expressions in equation (3) over all the subjects and the resulting log-likelihood is then maximized to estimate the latent-scale mean model parameters ($\boldsymbol{\beta^*}$).

3 Estimation of overall exposure effects in Tobit models

3.1 Direct-marginalization approach

In the *direct-marginalization* approach to the Tobit regression model, the marginal expectation of the natural scale response variable is linked to covariates and exposure status using a standard generalized linear model with identity link function

$$E(Y_i)=g(\mu_i)=\mu_i=\boldsymbol{X_i'\beta} \tag{5}$$

where $\boldsymbol{\beta}$ is the marginalized fixed effect regression coefficient vector of the covariate vector $\boldsymbol{X_i}$ on the observed outcome scale, leading to natural population-averaged parameter interpretations, i.e. coefficient for exposure represents the amount by which the overall mean μ_i changes from the unexposed group to the exposed group after adjusting for other covariates. We use the maximum likelihood techniques to estimate the $\boldsymbol{\beta}$ parameters, utilizing the latent dependent Tobit likelihood in equation (3), and the relationship of the specified marginal mean $\mu_i=\boldsymbol{X_i'\beta}$ with the Tobit latent mean μ_i^* from equation (4). The solution for the latent dependent variable mean, $\mu_i^*=\Delta^{-1}(\mu_i)$, can be found using standard numerical approaches such as the Newton–Raphson method. With this reparameterization, marginalized parameter estimation can be implemented using a variety of available techniques, including the EM algorithm, or any number of nonlinear optimization algorithms (such as quasi-Newton).

3.2 APV approach

The *APV* approach involves the calculation of model-predicted responses across different exposure statuses for each subject using the latent Tobit model specification in formula (1) and the marginal

mean transformation in formulae (4). For each person, predictions can be made for each exposure status regardless of the person's actual observed exposure status while fixing other covariates at the person's observed values. We note that, since each person in truth is either exposed or not exposed, one of these two predicted values represents a counterfactual response. Overall exposure effects for the Tobit regression model are represented on the natural scale as the average difference in expected response variables over the covariate distribution for a chosen reference population (for example, the exposed group for a binary predictor), estimated by

$$\delta = \int [E(Y \mid T = 1, \boldsymbol{w}) - E(Y \mid T = 0, \ \boldsymbol{w})] dF(\boldsymbol{w}) \tag{6}$$

where $F(\boldsymbol{w})$ is the joint distribution for the covariate vector $\boldsymbol{w}$ in the reference population and the integral is over the covariate space of $\boldsymbol{w}$. When the covariates ($\boldsymbol{w}$) are discrete, the integral in (6) may be written as a sum. For either categorical or continuous covariates, the true distribution function for $\boldsymbol{w}$ is typically unknown, so a common approach is to use the empirical distribution function by summing over the observed multivariate covariate values in the reference group (denoted by G with sample size n_G). Thus, our estimate of the average difference for the Tobit model becomes

$$\hat{\delta} = \frac{1}{n_G} \sum_{i \in G} [E(Y_i \mid t_i = 1, \boldsymbol{w_i}) - E(Y_i \mid t_i = 0, \boldsymbol{w_i})] \tag{7}$$

For the Tobit regression model, the expected value μ_i ($E(Y_i \mid t_i, \ \boldsymbol{w_i})$) in formula (7) for an individual with observed covariate values $\boldsymbol{w_i}$ and given exposure status t_i can be expressed as a function of μ_i^*, σ, l, and u as in formulae (4) in which μ_i^* and σ are estimable by plugging in estimated coefficients following the fit of the model (1) and (2) to the whole sample. An estimate of the mean difference δ is then obtained by averaging the predicted μ_i difference ($t_i = 1$ vs. 0) over the empirical distribution of the covariates ($\boldsymbol{w}$) in the reference group G.

3.3 Contrasting the two approaches

The *direct-marginalization* approach models the truncated normally distributed response variable, and thus assumes a homogeneous exposure effect on the observed, natural scale of the dependent variable. The *APV* approach instead models the latent dependent variable and thus assumes a homogeneous exposure effect on the latent response scale. Likelihood-based inference in both approaches will allow us to use fit criteria such as the Akaike information criterion (AIC)[29] and the Bayesian information criterion (BIC)[30] to help guide the choice of approach in real life analysis (smaller AIC/BIC indicates better fit). In the present paper, estimation for both approaches was implemented using quasi-Newton nonlinear optimization, implemented in SAS 9.3 IML (SAS Institute, Cary, NC, USA).

4 Simulation study

In this section, we use simulation studies to further investigate the properties of the proposed methods. Our primary goals were to determine the validity and robustness of both approaches under a variety of scenarios. Robustness is important as there are two primary underlying data mechanisms considered. We examined the robustness of the *direct-marginalization* approach when exposure effects were in fact generated as homogeneous on the latent response scale (true latent

Tobit regression model), and we also examined robustness of the *APV* approach when effects were generated as homogeneous for the marginal mean (true marginal Tobit regression model).

4.1 Simulation study design and methods

In our first simulation study, we assumed the true latent Tobit regression model for a binary exposure. We considered the cases of no covariates, a continuous and a binary covariate including scenarios across different exposure effects and covariate effects. The regression coefficient values used in the simulation study for each of these scenarios are given in Table S1 of the Supporting Information. For the true latent Tobit regression model (1), the intercept (β_0^*) was set to 24 and the standard deviation of the error (σ) was set to 4.5 for all scenarios. We also considered different situations with regard to covariate balance as before.[24] Simply to say, we included a balanced case (''B'' in Table S1) where each exposure group had a 50% frequency of $w = 1$ for binary covariate case and in the continuous case, the expected value of the covariate was 60 for both exposure groups. We also included two unbalanced cases. In one case, the imbalance 'favored' the exposure (''E'' in Table S1); the frequencies of $w = 1$ were 90% in the exposed group and 10% in the unexposed group, and in the continuous covariate case, the expected values of the covariate were 60 in the exposed group and 18 in the unexposed group. In the other unbalanced case, the imbalance favored non-exposure (''U'' in Table S1); that is, the above proportions/means were used with the groups switched. For the Tobit regression model, the boundary values l and u were set to 0 and 30 to generate single upper boundary data (latent dependent variable will not reach lower boundary 0 with specified parameter coefficients) or set to 21 and 30 to generate double boundary data.

For each scenario and type of covariate balance, 1000 simulated datasets were generated. We used sample sizes of 200 (100 per exposure group) and 2000 (1000 per exposure group) and the model variables were generated using pseudorandom number generators in SAS/IML (SAS, Version 9.3). For the binary covariate and exposure indicator, the randomization was constrained to assure the targeted proportion. In the continuous case, the covariate was generated independently from a normal distribution with standard deviation of 10 and expected values specified above. The latent response variables were then generated independently according to the Tobit regression model (1) given the individual exposure and covariate values and then truncated with corresponding boundaries l and u. In the second simulation study, we also generated data under a true marginal Tobit regression model by assuming homogeneous exposure effects for the marginal mean. In this case, exposure and covariate values determined the marginal mean, and the latent dependent variable mean was calculated from nonlinear equation (4) using the Newton-Raphson method. Then, the latent dependent variable was generated and truncated to generate the response variable.

For each dataset, both *direct-marginalization* and *APV* methods were used to estimate the difference in overall means for the exposed versus unexposed groups, and to construct a 95% confidence interval for the difference. For the *APV* approach, estimators were calculated by summing over the empirical distribution function of the covariate from the exposed (reference) group. The estimated difference in overall means from the AIC selected model among these two approaches for each simulated data set was also summarized. From the simulations, we calculated the average estimated mean difference (MD); the average percent error (PE = 100 × (Average Estimated MD – True MD)/True MD); the standard deviation (SD) of the estimated MD; and the coverage probability (CP, percent of simulated datasets for which 95% confidence interval for MD covered the true value). In addition, we also summarized the latent Tobit exposure coefficient estimates and the exposure coefficient estimates from the linear regression model ignoring the censored structure in our simulated data.

4.2 Simulation study results

We focus on results for the continuous covariate case; results for the binary case are similar and are therefore not presented. Table 1 gives the results in the case of $n = 100$ per group on single boundary data generated from the true latent Tobit regression model. When there are no exposure effects, both marginalized and *APV* approaches work well with low bias and Type I error within 2% of the nominal level for no-covariate, balanced, or unbalanced cases (scenarios 1, 3, 4, 7, 8, 11, and 12). We see over the scenarios without covariates (scenario 2) or with balanced covariates (scenarios 5 and 6) assuming non-zero exposure effects, that both *direct-marginalization* and *APV* approaches produce a small bias in estimating the mean difference (average PE less than 1.5%), and the coverage probabilities of 95% confidence intervals are within 3.5% of the nominal level. In the unbalance case, the average PE is still less than 2.5% for *APV* approach. However, the *direct-marginalization* approach has relative biases of up to 25% in unbalanced situations when substantial covariate effects exist (scenario 10 favoring exposed). The SE is also underestimated in this scenario and coverage of the 95% confidence interval can be as low as 82%. Generally, the AIC selected approach produces an unbiased estimate of the overall exposure effect, but for scenario 10 in unbalanced case, the relative bias is 6.3% and the coverage probability of the 95% confidence interval is less than the nominal level. Table S2 in the Supporting Material shows the results in the case of $n = 100$ per group on the double boundary data. Similarly, the *APV* approach provides relatively low bias for all scenarios (less than 2.5%), and the *direct-marginalization* approach produces low bias for no-covariate and balanced-covariate scenarios, but is biased for true latent Tobit structures in the unbalanced case when strong covariate effect and non-zero exposure effect exist. The bias may also be corrected with the AIC selected approach. When the sample size per group is increased to 1000, the *APV* approach shows even lower bias (less than 1.5%) and good coverage (within 1.5% of the nominal level) for all scenarios. However, the *direct-marginalization* approach continues to show relative biases of up to 25% for the unbalanced case with medium covariate effects under a true latent Tobit model. In addition, the *APV* approach provides unbiased latent Tobit exposure coefficient estimates but overestimates the overall exposure effect on the original outcome scale, while the exposure effect estimates from the linear regression model are biased in the unbalanced case with medium covariate effect and non-zero exposure effect (scenario 10 and 14 in Table 1 and Table S2).

We also considered analogous true marginal Tobit regression models assuming homogeneous exposure effects for the marginal mean. In this case, the relative performance of the *APV* and marginalized methods are essentially reversed from the previous results in which the *direct-marginalization* approach is unbiased and the *APV* approach has biases of up to 14% in unbalanced situations with non-zero exposure effect and medium covariate effect (Table 2 and Table S3 in the supporting information). Although the *direct-marginalization* approach generated unbiased parameter estimates, in some simulation scenarios (e.g. scenario 10 in Table 2 and Table S3), the variance of the overall exposure effects generated from the Hessian matrix of the likelihood function was unstable for some simulated data sets and the coverage probability appears low with $n = 100$ per group (around 90%). When the sample size per group is increased to 1000, the coverage probability is increased to approximately the nominal level (data not shown). In this setting, AIC selection also improves estimation of the mean difference compared against the mis-specified *APV* approach (average PE $< 10\%$ for $n = 100$ per group and $< 5\%$ for $n = 1000$ per group). As expected, basic linear regression provided much poorer fit than the Tobit models, e.g. for scenario 10 in Table 2, and the mean AICs from 1000 simulated data sets are 1050.5 for the linear regression model and 869.5 for the marginal Tobit model using the *direct-marginalization* approach.

Table 1. Simulation statistics for the estimated overall mean difference with the *direct-marginalization*, *APV* methods, and AIC selected approach and the exposure coefficients on single boundary data generated from the true latent Tobit regression model without covariates or with one continuous covariate, $n = 100$ per group with 1000 simulations per scenario.

				Direct-marginalization approach				*APV* approach				AIC selected approach				Estimated latent Tobit exposure coefficient	Estimated linear regression exposure coefficient
Scenario	Balance[a]	Cov. Eff.[b]	True MD	Ave Est MD	Ave PE (%)	SD of Est MD	CP (%)	Ave Est MD	Ave PE (%)	SD of Est MD	CP (%)	Ave Est MD	Ave PE (%)	SD of Est MD	CP (%)		
1	N	–	0	−0.01	–	0.59	94.8	−0.01	–	0.59	94.9	−0.01	–	0.59	94.7	−0.01	−0.01
2		–	2.73	2.71	−1.0	0.51	96.1	2.71	−0.8	0.51	96.2	2.71	−0.8	0.51	96.3	3.27	2.71
3	B	S	0	0.02	–	0.57	94.0	0.03	–	0.57	94.2	0.02	–	0.57	94.1	0.03	0.03
4		M	0	−0.01	–	0.49	95.2	−0.01	–	0.50	95.2	−0.01	–	0.50	95.1	−0.01	−0.01
5		S	2.74	2.76	0.4	0.52	93.5	2.76	0.5	0.52	93.5	2.76	0.7	0.52	93.5	3.63	2.76
6		M	2.76	2.76	−0.2	0.81	91.8	2.75	−0.6	0.41	94.6	2.75	−0.6	0.41	92.4	4.78	2.74
7	E	S	0	−0.03	–	1.30	95.7	0	–	1.26	95.5	−0.02	–	1.29	95.7	−0.05	−0.03
8		M	0	0.07	–	1.26	93.7	0.10	–	1.15	93.1	0.05	–	1.22	92.7	0.04	0.05
9		S	2.74	2.85	3.9	1.16	94.2	2.77	1.0	1.33	92.5	2.79	1.7	1.30	92.3	3.57	2.83
10		M	2.76	3.46	25.2	1.06	82.8	2.83	2.3	1.19	92.5	2.93	6.3	1.16	86.2	4.79	3.14
11	U	S	0	0.07	–	1.34	94.0	0.14	–	1.32	94.4	0.10	–	1.34	93.9	0.11	0.08
12		M	0	−0.03	–	1.27	94.8	0.04	–	1.28	93.8	0.03	–	1.30	94.3	−0.01	−0.01
13		S	2.75	2.71	−1.4	1.13	95.5	2.80	1.7	1.27	95.2	2.79	1.6	1.25	94.6	3.42	2.73
14		M	2.76	2.45	−11.2	1.06	94.3	2.76	0.0	1.35	94.9	2.71	−2.0	1.28	93.8	3.55	2.53

MD: mean difference; SD: standard deviation; PE: percent error; CP: coverage probability.

[a]N, no covariates; B, balanced covariate; E, unbalanced favoring exposed; U, unbalanced favoring unexposed.

[b]Covariate Effect: S, small effect; M, medium effect.

Table 2. Simulation statistics for the estimated overall mean difference with the *direct-marginalization*, *APV* methods, and AIC selected approach and the exposure coefficients on single boundary data generated from the true marginal Tobit regression model without covariates or with one continuous covariate, $n = 100$ per group with 1000 simulations per scenario.

				Direct-marginalization approach				*APV* approach				AIC selected approach				Estimated latent Tobit exposure coefficient	Estimated linear regression exposure coefficient
Scenario	Balance[a]	Cov. Eff[b]	True MD	Ave Est MD	Ave PE (%)	SD of Est MD	CP (%)	Ave Est MD	Ave PE (%)	SD of Est MD	CP (%)	Ave Est MD	Ave PE (%)	SD of Est MD	CP (%)		
1	N	–	0	0.01	–	0.58	94.7	0.01	–	0.58	94.8	0.01	–	0.58	94.7	0.01	0.01
2		–	2.75	2.73	−0.8	0.53	94.4	2.73	−0.6	0.53	94.5	2.73	−0.6	0.53	94.3	3.36	2.73
3	B	S	0	0.02	–	0.57	94.1	0.02	–	0.57	94.1	0.02	–	0.57	94.1	0.02	0.01
4		M	0	0.03	–	0.52	94.1	0.03	–	0.53	93.5	0.03	–	0.53	93.7	0.04	0.03
5		S	2.75	2.74	−0.4	0.50	94.0	2.74	−0.3	0.50	94.3	2.75	−0.2	0.50	94.2	3.59	2.74
6		M	2.75	2.75	−0.2	0.46	92.3	2.76	0.2	0.42	95.0	2.76	0.3	0.42	92.9	4.45	2.75
7	E	S	0	0.08	–	1.30	95.5	0.11	–	1.27	94.7	0.09	–	1.30	94.9	0.08	0.08
8		M	0	−0.11	–	1.29	94.2	−0.05	–	1.18	93.7	−0.10	–	1.27	93.6	−0.14	−0.11
9		S	2.75	2.75	0.1	1.16	93.7	2.70	−1.7	1.34	92.4	2.71	−1.3	1.32	92.4	3.50	2.76
10		M	2.75	2.79	1.5	1.14	89.6	2.36	−14.0	1.17	89.9	2.48	−9.9	1.13	87.9	3.85	2.77
11	U	S	0	0.01	–	1.32	94.5	0.07	–	1.30	95.1	0.04	–	1.33	94.7	0.04	0.01
12		M	0	−0.02	–	1.29	93.5	0.05	–	1.30	94.0	0.03	–	1.31	93.0	0.01	−0.02
13		S	2.75	2.70	−1.8	1.12	95.6	2.78	1.1	1.26	95.6	2.77	0.8	1.23	95.3	3.44	2.72
14		M	2.75	2.70	−1.6	1.06	94.8	3.00	9.0	1.34	94.4	2.94	6.9	1.26	93.6	3.82	2.75

MD: mean difference; SD: standard deviation; PE: percent error; CP: coverage probability.

[a]N, no covariates; B, balanced covariate; E, unbalanced favoring exposed; U, unbalanced favoring unexposed.

[b]Covariate Effect: S, small effect; M, medium effect.

5 Application to the Genoa cohort study data

Phase II GENOA participants underwent a neurocognitive testing battery to assess several domains of cognitive function in an ancillary study conducted between August 2001 and May 2006. We focus here on the MMSE score which ranges from 0 to 30 with our primary objective being to compare this lower and upper bounded cognitive measure for obese (BMI ≥ 30) versus non-obese (BMI < 30) non-Hispanic white participants. For illustrative purposes, we included age, gender and smoking as baseline covariates in the analysis and assumed that study participants were independent. The GENOA phase II dataset used for the analysis included 850 subjects with complete data for all the variables specified in the models, and approximately one-half of the participants ($n = 414$) were classified as obese according their BMI. To compare alternative approaches, we fit the following models: (1) a linear regression model ignoring the boundaries of the data; (2) a latent Tobit regression model, with additional *APV* marginalization; and (3) a marginal Tobit model using the *direct-marginalization* approach. According to the AIC and BIC fit criteria (Table 3), both Tobit models are superior to the linear regression model and the latent Tobit model provides a slightly better fit to the data than the marginal Tobit model. Figure 1 shows the predicted frequencies for the linear regression model and two Tobit regression models against the histogram of observed frequencies for the MMSE score. Consistent with the AIC and BIC values, the linear regression model provides a poor fit (Figure 1a) and the predicted frequencies for the latent Tobit model fit slightly better than the marginal Tobit model (Figure 1b). The poor fit for the linear regression model is expected due to the fact that this model fails to take into account the right censoring of our MMSE scores. Although the latent Tobit model and the marginal Tobit model provide satisfactory fit, the predicted frequencies from both Tobit models are somehow underestimated in two categories with the largest MMSE score (28.7–29.9 and 30) and overestimated in several other categories (25.0–26.2, 26.3–27.4, and 27.5–28.6) as compared to the observed frequencies, which suggests that other un-included covariates (e.g. education or physical activity) may significantly predict the MMSE score in the GENOA cohort study data.

Table 3 shows inferential results for cognitive differences between obese and non-obese participants using the MMSE score across different models. The table includes the estimated mean difference (mean for obese participants minus mean for non-obese participants), the related standard error (SE) and the Wald test p-value of the null hypothesis of the mean difference equal to 0. The linear regression model using the ordinary least squares method provides an estimated mean

Table 3. Model selection criteria and estimated difference of the mean MMSE score from alternative models for the obese and non-obese non-Hispanic Whites controlling for age, gender and smoking in the GENOA Study Phase II ($N = 850$).

	Fit criteria		Mean difference		
	AIC	BIC	Estimate	SE	*P*
Linear regression model estimate	3020.1	3048.6	−0.098	0.098[a]	0.317
Latent Tobit regression model	2669.6	2694.7			
Tobit model exposure coefficient			−0.267	0.165	0.106
APV approach estimate			−0.151	0.094	0.108
Marginal Tobit regression model	2679.8	2704.8			
Direct-marginalization approach estimate			−0.117	0.096	0.223

[a]Robust (heteroscedasticity-consistent) standard error estimate from the linear regression model.

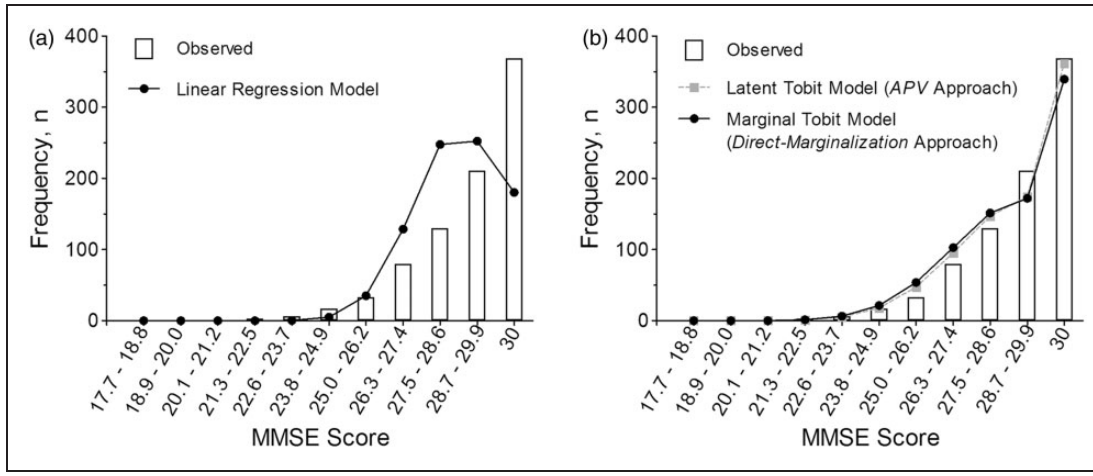

Figure 1. Observed and predicted frequencies for each MMSE category in GENOA Study Phase II. Predicted values are from the linear regression model (panel (a), solid line, black "•"), the latent Tobit regression model (panel (b), dashed line, gray "■"), and the marginal Tobit regression model (panel (b), solid line, black "•") respectively.

difference of −0.098 with robust standard error 0.098, but had a poor fit to the data as described above. The AIC selected latent Tobit model gave the overall exposure effect estimate of −0.151 from the *APV* approach, indicating an increase of 54% over the linear regression results. Similar to the *APV* approach, the marginal Tobit model using *direct-marginalization* gave the estimate as −0.117, an increase of 19% over linear regression. Both of these estimates suggest that obese participants have lower mean MMSE scores than non-obese participants and basic linear regression may underestimate this effect; however, this difference in not statistically significant by either method. Of note, the exposure effect estimate on the latent dependent variable from the Tobit regression model was −0.267, 77% higher than the overall exposure effect estimate of −0.151 from the *APV* approach, which is an overestimation of the actual mean MMSE score difference between the two exposure groups. In the light of our simulation results, the minor difference in estimates from the *APV* and *direct-marginalization* approaches can be explained given the slightly imbalanced distributions of gender and age between obese and non-obese participants (female proportion 59.4% vs. 57.2% and age mean (SD): 60.1 (10.2) vs. 58.5 (10.0)).

6 Discussion

In this article, we have studied the use of Tobit regression models for comparing overall response means between groups while controlling for baseline covariates. The Tobit regression models are appealing because of their ability to account for censoring in the dependent variable, relative to the linear regression model, allowing them to provide a good fit to some data with measurement limit in medical research. Overall exposure effects are estimable for the Tobit regression model using a *direct-marginalization* or *APV* approach as described in this paper. Both methods have been implemented in a SAS Macro, which is available for downloading from the first author's webpage, http://www.umc.edu/faculty/wei_wang/.

The first proposed method, the *direct-marginalization* approach, models the truncated dependent variable mean directly, and the exposure coefficient provides the estimated overall exposure effects. A similar approach, as we discussed in the introduction, has been used by Long et al. in ZIP regression models.[17] An advantage of our *direct-marginalization* approach is that numerical root-finding methods are introduced and thus allows nonlinear relationship between latent dependent variable mean and marginal mean for likelihood specification and parameter estimation. Similar approaches have been proposed for marginalizing mean parameters over random effects distribution in mixed effects models[20–22]; our article extends these to nonlinear Tobit specifications and is the first article we are aware of that proposes to marginalize two components of a mixture model to estimate the overall exposure effects. Our second proposed method, the *APV* approach, involves the comparison of model-predicted response values for each individual under different exposure statuses and has relationships to other post-estimation effect constructions.[25–28]

Although both approaches can provide overall exposure effect estimators, they actually use different underlying model settings. The *direct-marginalization* approach assumes homogeneous effects for the marginal mean, whereas the *APV* approach assumes homogeneous effects on the latent response scale. An interesting finding of our simulation studies is that, both approaches appear to provide valid, robust inferences in the case of a null exposure effect, or in general exposure effect cases with balanced covariates or no covariates. However, in the case of an unbalanced covariate, both methods appear less robust (providing potentially biased estimates) when assuming the incorrect model, particularly when the covariate has a large effect on the outcome. Considering that the *APV* approach assumes homogeneous exposure effect on the latent response scale and the required overall exposure effect estimation in this approach is determined by the covariate distribution in the chosen reference population (see formula (7)), it is not surprising that the *direct-marginalization* approach and the *APV* approach provide inconsistent overall exposure effect estimates when unbalanced covariates with substantial effects exist. Our simulation results support the use of AIC or BIC to determine the model setting, and the bias on the overall exposure effect estimates from AIC selected approach is lower than that from the mis-specified model underlying two different model settings. Specifically, for an unbalanced covariate favoring the exposed group (scenario 10 in Table 2) assuming a true latent Tobit regression model, in 60.3% of data sets, the *APV* approach had lower AIC than the *direct-marginalization* approach. When the sample size was increased to 1000 per group, the percentage was 77.0%. Similarly, in the comparable scenario assuming a true marginal Tobit regression model, the *direct-marginalization* approach had lower AIC than the *APV* approach in 58.2% of data sets for sample size of 100 and 79.3% of data sets for a sample size of 1000 per group. In addition, the marginal exposure effect from the linear regression model can be biased when the data are generated from the latent Tobit model, and the fit for this model is much worse than the Tobit models due to ignoring the censored structure.

Both the *direct-marginalization* approach and *APV* approach use the Tobit regression model likelihood for parameter estimation. The overall exposure effects estimation requires additional post-modeling computation for the *APV* approach, while related additional computation is directly incorporated in model estimation for the *direct-marginalization* approach. The *direct-marginalization* approach may have difficulty in converging when a large proportion of the response reaches the boundary, possibly due to extensive computations to specify the likelihood in this situation. The variance of the estimated exposure effect can be obtained directly from the Hessian matrix of likelihood function for the *direct-marginalization* approach and the delta method for the *APV* approach. Alternatively, variance estimates can also be obtained via bootstrap resampling[31] which allows the computation of confidence intervals without the requirement of an asymptotic normality assumption for the estimator.

In conclusion, using either a *direct-marginalization* or an *APV* approach provides the overall effect estimates with more natural interpretations and more direct connections to the observed data for diagnostic checking. A key assumption distinguishing these approaches is whether the exposure effects are assumed to operate on the observed (marginal) data scale, or the latent Tobit scale. For real world applied analyses, we recommend using fit criteria (AIC or BIC) to determine which regression model, marginal versus latent Tobit, appears to have greater support, with the note that both approaches appear robust in balanced covariate settings. Future research is needed to extend these methods to Tobit regression models with multi-level correlated data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work is supported in part by the National Institute on Aging, National Institutes of Health Research [grant number R01AG045255 (B.G. Windham)].

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website.

References

1. Fox ER, Musani SK, Bidulescu A, et al. Relation of obesity to circulating B-type natriuretic peptide concentrations in blacks: The Jackson Heart Study. *Circulation* 2011; **124**: 1021–1027.
2. Wang L, Zhang Z, McArdle JJ, et al. Investigating ceiling effects in longitudinal data analysis. *Multivariate Behav Res* 2009; **43**: 476–496.
3. Tobin J. Estimation of relationships for limited dependent variables. *Econometrica* 1958; **26**: 24–36.
4. Wang TJ, Larson MG, Keyes MJ, et al. Association of plasma natriuretic peptide levels with metabolic risk factors in ambulatory individuals. *Circulation* 2007; **115**: 1345–1353.
5. Moulton LH and Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* 1995; **51**: 1570–1578.
6. Moulton LH, Curriero FC and Barroso PF. Mixture models for quantitative HIV RNA data. *Stat Methods Med Res* 2002; **11**: 317–325.
7. Elkins JS, Johnston SC, Ziv E, et al. Methylenetetrahydrofolate reductase C677T polymorphism and cognitive function in older women. *Am J Epidemiol* 2007; **166**: 672–678.
8. Giang LM, Weiner DE, Agganis BT, et al. Cognitive function and dialysis adequacy: no clear relationship. *Am J Nephrol* 2011; **33**: 33–38.
9. FBPP Investigators: Multi-center genetic study of hypertension: The Family Blood Pressure Program (FBPP). *Hypertension* 2002; **39**: 3–9.
10. Twisk J and Rijmen F. Longitudinal Tobit regression: A new approach to analyze outcome variables with floor or ceiling effects. *J Clin Epidemiol* 2009; **62**: 953–958.
11. Fu P, Hughes J, Zeng G, et al. A comparative investigation of methods for longitudinal data with limits of detection though a case study. *Stat Methods Med Res* 2012; DOI: 10.1177/0962280212444800.
12. Sattar A, Weissfeld LA and Molenberghs G. Analysis of non-ignorable missing and left-censored longitudinal data using a weighted random effects tobit model. *Stat Med* 2011; **30**: 3167–3180.
13. Dagne GA and Huang Y. Bayesian semiparametric mixture Tobit models with left censoring, skewness, and covariate measurement errors. *Stat Med* 2013; **32**: 3881–3898.
14. Chu H, Kensler TW and Muñoz A. Assessing the effect of interventions in the context of mixture distributions with detection limits. *Stat Med* 2005; **24**: 2053–2067.
15. McDonald JF and Moffitt RA. The uses of tobit analysis. *Rev Econ Stat* 1980; **62**: 318–321.
16. Kang J. The usefulness and uselessness of the decomposition of Tobit coefficients. *Sociol Method Res* 2007; **35**: 572–582.
17. Long DL, Preisser JS, Herring AH, et al. A marginalized zero-inflated Poisson regression model with overall exposure effects. *Stat Med* 2014; **33**: 5151–5165.
18. Süli E and Mayers D. *An introduction to numerical analysis*. Cambridge: Cambridge University Press, 2003.
19. Brent RP. *Algorithms for minimization without derivatives*. Englewood Cliffs, NJ: Prentice Hall, 1973.
20. Lee K, Joo Y, Song JJ, et al. Analysis of longitudinal zero-inflated count data using marginalized models. *Comput Stat Data Anal* 2011; **55**: 824–837.
21. Iddi S and Molenberghs G. A combined overdispersed and marginalized multilevel model. *Comput Stat Data Anal* 2012; **56**: 1944–1951.

22. Heagerty PJ. Marginalized specified logistic-normal models for longitudinal binary data. *Biometrics* 1999; **55**: 688–698.
23. Pawitan Y. *In all likelihood: Statistical modelling and inference using likelihood*. Oxford: Oxford University, 2001.
24. Albert JM, Wang W and Nelson S. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Stat Methods Med Res* 2011; **23**: 257–278.
25. Greenland S. Model-based estimation of relative risks and other epidemiological measures in studies of common outcomes and in case-control studies. *Am J Epidemiol* 2001; **160**: 301–305.
26. Localio AR, Margolis DJ and Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *J Clin Epidemiol* 2007; **60**: 874–882.
27. Bender R, Kuss O, Hildebrandt M, et al. Estimating adjusted NNT measures in logistic regression analysis. *Stat Med* 2007; **26**: 5586–5595.
28. Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *J Clin Epidemiol* 2010; **63**: 2–6.
29. Akaike H. A new look at the statistical model identification. *IEEE T Automat Contr* 1974; **19**: 716–723.
30. Schwarz GE. Estimating the dimension of a model. *Ann Stat* 1978; **6**: 461–464.
31. Efron B and Tibshirani R. *An introduction to the bootstrap*. New York: Chapman and Hall, 1993.