ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/medsci.2017-54827

Usage of Heckman Sample Selection Model in Health Studies: An Application of Prostate Cancer Patients

Heckman Seçim Modellerinin Sağlık Alanında Kullanımı: Prostat Kanserli Hastalar Üzerine Bir Uygulama

Özge PASİN,[a]
Handan ANKARALI,[b]
Levent YEŞİLYURT[c]

[a]Department of Biostatistics,
İstanbul University
İstanbul Faculty of Medicine,
İstanbul
[b]Department of Biostatistics,
Düzce University Faculty of Medicine,
Düzce
[c]Clinic of Radiology,
Palandöken Public Hospital,
Erzurum

Geliş Tarihi/*Received:* 16.01.2017
Kabul Tarihi/*Accepted:* 14.08.2017

Yazışma Adresi/*Correspondence:*
Özge PASİN
İstanbul University
İstanbul Faculty of Medicine,
Department of Biostatistics, İstanbul,
TURKEY/TÜRKİYE
ozgepasin90@yahoo.com.tr

ABSTRACT Objective: The objective of this study is to introduce theoretical characteristics of Heckman sample selection regression model,to clarify when it is required to be used and to show its usage in health area. **Material and Methods:** Heckman selection model, which is an appropriate tool for addressing the sample selection bias, was used for determining the risk factors of Gleason score for three datasets which have 2000 observations obtained from simulation study. The data was about the prostate cancer patients. In the data, there were benign tumor structures or stage 1 tumors in the study besides malign tumors but Gleason score was not calculated when tumor is benign. So sample selection bias is a matter for nonresponse Gleason score answers in our data. Also the model was performed for the datasets that have 0.30; 0.50 and 0.70 of censored proportion (nonresponse Gleason score) individual number by considering actual structures between variables with the help of simulation. **Results:** There was no significant relationship between Gleason score and smoking, family history. But we found that age, PSA(prostate specific antigen) and weight variables have significant relationship with Gleason score (p<0.001). Adjusted standard error value was the highest in dataset with 0.50 censored proportion, and the lowest in data set with highest censor proportion (0.70). Among models the lowest log likelihood value has been determined in model with 0.50 censor proportion. **Conclusion:** When there is sample selection bias on dependent variable, Heckman sample selection regression model can be suggested. In addition, researchers can have more accurate results by increasing the use of these models in health sciences.

Keywords: Models, statistical; neoplasm grading; prostatic neoplasms; selection bias

ÖZET Amaç: Bu çalışmanın amacı, Heckman seçim regresyon modelinin teorik özelliklerini tanıtmak, hangi durumlarda kullanılması gerektiğini ve sağlık alanında kullanımını göstermektir. **Gereç ve Yöntemler:** Çalışmada örnek seçim yanlılığını gideren uygun bir model olan Heckman seçim modeli, Gleason skoruna etki eden risk faktörlerini araştırmak amacıyla simülasyon çalışmasından elde edilen 2000 gözleme sahip üç veri seti için kullanılmıştır. Veri setleri prostat kanserli hastaları içermektedir. Gleason skoru prostat kanseri hastalarında tümörün çeşidi malign olduğunda hesaplanan bir skor olmasına rağmen, çalışmamızda tümor yapısı olarak iyi huylu veya 1. evre tümörler de mevcut olduğundan, Gleason skoru hesaplanamayan birçok kişi sebebiyle örnek seçim yanlılığı söz konusudur. Model, gerçek yapıları dikkate alarak simülasyon yoluyla elde edilen sansür oranları (Gleason skoru hesaplanamayan) sırasıyla 0.30; 0.50 ve 0.70 olan üç farklı veri seti için kurulmuştur. **Bulgular:** Gleason skoru ile sigara içme durumu ve aile öyküsü değişkenleri arasında istatistiksel olarak anlamlı bir ilişki gözlenmemiştir. Ancak, Gleason skoru ile yaş, PSA ve ağırlık arasında anlamlı bir ilişki gözlenmiştir (p<0.001). Düzeltilmiş standart hata değeri en yüksek 0.50 sansür oranına sahip veri setinde ve en düşük değer ise sansür oranı 0.70 olan veri setinde elde edilmiştir. En düşük log olabilirlik değerine sahip model ise 0.50 sansür oranına sahip olan Heckman modelinde elde edilmiştir. **Sonuç:** Bağımlı değişken üzerinde seçim yanlılığı varlığında, Heckman seçim regresyon modellerinin kullanımı düşünülebilir. Ayrıca araştırmacılar, bu modellerin sağlık alanında kullanımını artırarak daha doğru sonuçlar elde edebilmektedir.

Anahtar Kelimeler: Modeller, istatistiksel; neoplazi derecelendirmesi; prostat neoplazileri; seçim yanlılığı

Regression models are one of the most used statistical models in applications frequently. A wide variety of regression models have been developed depending on the number and the type of dependent and independent variables in the models. One of the models in these is tobit model. This model was developed in 1958 and it is different than other regression models in terms of obtaining dependent variable values. The model has entered in application area in recent years.[1,2] Tobit model has been derived from tobin and probit names. It is an extension of probit models. Tobit analysis is used in cases when some of dependent variable values can not be observed in correspondance to known values of independent variable contrary to linear regression models. In these models dependent variable value is sometimes censor from top sometimes from bottom part. When data censoring occurs, estimators of least squares linear models have been proved to be biased and inconsistent with simulation studies in many researches. So, the continuous dependent variable is censored at a specific value. But there is no limitation on the way of obtaining of independent variables. Tobit models can be used when dependent variable can only be valued at specific ranges (for example data limited from top and bottom like success notes, scale points) and when subjects are not related to objective of the study (for example in a study where risk factors effecting the amount of smoking are researched, the non-smoking individuals consuming of cigarette will be zero). Also in this circumstance, the Heckman sample selection model can be used that removes subjects from study. This model is a type of tobit models.[1-3] Heckman models concentrates on "incidental truncation" of the dependent variable. It address the presence of sample selectivity bias.

In many researches, selections of the subjects and the randomization of the subjects in a study is an important problem. If the subjects are not selected by a randomization process, it is inevitable that a selection bias occurs in the research.[4] Heckman model is used for correcting selection bias and adjusting bias that may occur from non-random sample selection bias statistically.[5,6] Therefore, before starting the model process of Heckman, researcher should consider whether there is a selection bias or not. If the zeros on the dependent variable are actual zeros which are observed, the data are treated as true zeros rather than missing values. So there is no selection bias here. But when the zeros are potential, the data are not treated as true zero, the data are censored. Researcher should use latent variable for this selection bias. Suppose, a sample contains a non-negligible proportion of subjects that who do not smoke. For these subjects, there is no information about the consuming amount of cigarette. So the corresponding observations cannot be used when the estimating of smoking amount. Thus there is non random sample to estimate dependent variable.

The Heckman models are often used in sociology and especially in economical researches. On the other hand, the use of this model in health researches is very rare.[7,8]

The objective of this study is to introduce the theoretical characteristics of Heckman sample selection regression model and to indicate how to use it in a health data set. With this model we want to determine the risk factors of Gleason scoring in the prostate cancer patients.

METHODS AND APPLICATION

HECKMAN SAMPLE SELECTION MODEL

In studies sample selection bias arises when the residual in the selection model (includes only response subjects) and the residual in the primary model (include all subjects) have correlation whenever the covariance of these residuals is not equal to zero. When the sample selection bias arises the Heckman model process should be started. In Heckman models, selection equation model is generated firstly and a probit model is established for observed predictors and likelihoods are estimated for each subjects. Following all these processes mills ratios, which is an statistical correction, are calculated. This ratio is calculated by using selection equation as independent variable for result variable considered in ordinary least squares (OLS) regression. Heckman model approaches selection bias as an omitted variable bias.

So as stated above, Heckman sample selection model performs estimation processes in two phases.

The primary equation in the model is defined as in the below,x

$$y_1 = x_1\beta_1 + u_1$$

x_1 is the dependent variable, β_1 is the coefficient of the model and u_1 is the error term.

There is a condition for the observation of dependent variable.

$$y_2=\begin{cases} 1 & if\ x\delta_2 + u_2 \geq 0 \\ 0 & otherwise \end{cases}$$

If $y_2 = 1$, y_1 can observed. The x values is always observed, regardless of the value.

In Heckman model, probit estimation method is applied for the whole dataset by using dummy variable.

By dummy variable being d_i;

$$d_i=\begin{cases} 1: if\ y_i > 0 \\ 0: if\ y_i = 0 \end{cases}$$

After dummy variable is defined probit (or logit) model is formed.

$d_i=x_i'\beta + \varepsilon_i,\ \varepsilon_i$ is error term and it disperses normally by its variance being σ^2, and average being zero. By using Probit (or logit) model $\frac{\hat{\beta}}{\sigma}$ is estimated.

In the second step, hazard function's estimation is calculated by using. $\frac{\hat{\beta}}{\sigma}$

So the model is,

$$E(y_1|x,y_2 = 1) = x_1\beta_1 + \Upsilon_1\lambda(x\delta_2)$$

$\lambda(x\delta_2)$ is the mills ratio of the observations. It is calculated as,

$$\lambda(x\delta_2) = \frac{\phi(x\delta_2)}{\Phi(x\delta_2)}$$

In the above equation $\emptyset$ is the standard normal probability density function and is the$\emptyset$ standard normal cumulative density function. So in Heckman model hazard (invers mill) function is defined as.

$$\lambda(x\delta_2) - \lambda(\mathrm{z}) - \frac{\phi(x\delta_2)}{\Phi(x\delta_2)}=\frac{\phi(z)}{\Phi(z)}; \qquad \mathrm{z}=\frac{y_i-x_i'\beta}{\sigma}.$$

Heckman model have corrected for sample selectivity by adding the model mills ratio.

After performing these steps, the correlation between error terms (rho coefficient (ρ)) obtained from both steps is tested whether it is equal to zero or not. If null hypothesis is rejected, that means $\rho\neq0$, Heckman selection model will gives more correct and effective results than standard regression models.[7-11]

An Application: Prostate cancer ve Gleason score

In our study, the risk factors effective on Gleason scores of patients with prostate cancer are investigated. The prostate cancer incidence is lower in Asian countries and is higher in Northern America countries. It is the most common type cancer in men in the United States of America and it accounts for 29% of all cancer cases. 95% of men having prostate cancer diagnosis are between 45-89 ages. Hormons, infections, diet, environmental factors and genetic predisposition plays important roles in disease etiology. Also in the literature, having a positive family history, being a member of black American race and excessive smoking are indicated as risk factors.

Clinicians can use Gleason scoring for histopathological rating system in prostate cancer patients. These score gives information about agre sivity, growth rate and dispersion degree of the tumor. Generally in performed studies when tumor type is malign, with regards to progress of the tumor, this score is considered. Thus when the tumor is malign, interpretation of the score becomes more of an issue. The pathologist determines the two most common degrees of difference in tissues obtained by biopsy and gives a total "Gleason score", which is the most frequent one. The Gleason score is a value between 2 and 10. Sum of the result being over 7 or any of the two scores being over 4 indicates a poor prognosis, being 10 indicates that the tumor is pretty agressive. Risk factors that effect rate of Gleason score can be psa (prostate specific antigen) density, histological extension of the tumor, obesity, age and also smoking, family history.[12,13]

In the study, the data was obtained from simulation by considering real structures and with the help of simulation, censored individual number ratio (non response answers in Gleason score) has been taken as 0.30, then this proportions has been taken as 0.50 and last this proportion has been taken as 0.70. The total observation number is 2000 for each dataset. The reason of this different censor occurs due to non-performance of Gleason score measurement in patients. We worked in three different censored proportions. Because we want to see the effects of censoring proportion in Heckman model for determining the risk factors. Thus by studying on the same data set, without changing the independent variable values, tumor type values has been changed and censor number has been increased or decreased. In selection phase of Heckman model, age, smoking, family history variables that may effect sample selection bias has been taken into model. In order to calculate Log likelihood value, no limitations has been made in iteration number. Stata 14 programme has been used in calculations.

RESULTS

Descriptive statistics was given Table 1 and Table 2 for the three different data set which has respectively censored data proportion of 0.30; 0.50; 0.70; regarding age, psa, weight, tumor type, smoking status and family history. The Gleason scores were between 2-8. For the data that have 0.30 and 0.50 censored proportion the Gleason scores were between 1-6, and for the last data set this range was between 1-8. When Table 1 is examined, nonresponders proportion in terms of Gleason score in the first data set was 30% , the proportion was observed as 50% in second data set and 70% for third data set. In all data sets the proportion of smokers was 51.3% (1027 observations) and approximately 50% (999 observations) of observations had family history in terms of prostate cancer (Tables 1, 2).

The Heckman sample selection model was performed for three data sets, all three models were concluded to be ssignificant.

In the first step, the selection model was created in the Heckman model. The model is a choice model that whether Gleason score was calculated or not. This model estimates the Gleason score presence status (absent/present). The selection model used binary outcome (non reponders, responders) and analysis were obtained by a probit model. The other step of Heckman model is regression equation. To express the model and understant it better, we can write two equations,

TABLE 1: Descriptive statistics of tumor types, smoking status and family history and number of patients for the three different dataset.

			Frequency	Percent
Proportion of censoring=0.30	Gleason score	Responders	1400	70
		Nonresponders	600	30
Proportion of censoring =0.50	Gleason score	Responders	1000	50
		Nonresponders	1000	50
Proportion of censoring =0.70	Gleason score	Responders	600	30
		Nonresponders	1400	70
For three data set	Smoking Status	No	973	48.7
		Yes	1027	51.3
	Family History	No	1001	50.1
		Yes	999	49.9

TABLE 2: Descriptive statistics of age, PSA and weight for the three different dataset.

Variable	N	Mean	Standard Deviation	Minimum	Maximum
Age (year)	2000	80.39	9.705	36	104
PSA	2000	14.02	7.630	2	29
Weight	2000	95.44	17.781	21.12	137.08

Regression equation

Gleason Score= $\beta_0+\beta_1\,\alpha ge+\beta_2 psa+\beta_3 smoking+\beta_4\ family\ history+\beta_5\ weight+u_1$

Gleason score is observed if

$\gamma_0+\gamma_1\ age+\gamma_2 smoking+\gamma_3 family>0$ selection model)

The mills ratio was obtained for the models and the values were given in (Table 3).

In table 4, the significance test results of the standard errors were given for the models. Adjusted standard error value (sigma) has been found highest in data set with benign ratio of 0.50 while it has the lowest value in data set with benign ratio of 0.70 (highest censor ratio). In order to decide which model is more fit, log likelihood values have been examined. The Log likelihood value close to zero proves that model is more fit. In established models the lowest log likelihood value was in the model with 0.50 censor proportion. Model closest to zero is the model with 0.70 censor proportion. Namely the best fitting model was, the model where Heckman sample selection model was used for the data set with the highest censor proportion. (Table 4).

TABLE 3: The mills ratio.

Censored proportion	Mills ratio
Proportion of benign tumor= 0.30	0,706
Proportion of benign tumor= 0.50	0,607
Proportion of benign tumor= 0.70	-0,639

Values of rho coefficients that examines relationships between errors are given in Table 5. This coefficient shows the correlation between the unobservables in the regression and unobservables in the selection model. When p values that test significance of Rho coefficients have been examined, for all three models rho coefficients have been concluded to be significant. This result has indicated that error value obtained in first step was in a mutual interaction with error value obtained from second step. Thereby for all three data sets using Heckman sample selection model has been concluded to be correct. When *Rho* coefficients have been examined differentness according to censor ratio has been observed. When censor proportion is 0.70 coefficient value is negative, while for other data sets this coefficient took positive value. Taking

TABLE 4: Descriptive statistics of age, PSA and weight for the three different dataset.

Censored Proportion	Valuec	%95 Confidence Interval		Log-Likelihood Value	p-values for model significance
Proportion of censoring= 0.30	0.824	0.780	0.870	-2134.924	<0.001
Proportion of censoring = 0.50	0.863	0.790	0.941	-2274.631	<0.001
Proportion of censoring = 0.70	0.795	0.690	0.915	-1681.88	<0.001

TABLE 5: Significance of Rho coefficient.

Censored proportion	Coefficient	Standard error	Chi-square value	p
Proportion of Benign tumor= 0.30	0.858	0.0503	32.74	<0.001
Proportion of Benign tumor= 0.50	0.704	0.0615	17.24	<0.001
Proportion of Benign tumor= 0.70	-0.804	0.055	9.54	<0.001

negative coefficient means that as unmeasurable variable value increases, possibility of being malign increases and Gleason score decreases. But in other two models (when censor ratio is 0.30 and 0.50) *rho* coefficient value has been close to each other and positive. Being positive means that, if there is a positive corelation between unobservable variable value and Gleason score, when unobservable variable value increases possibility of being malign for the tumor type also increases (Table 5). So as a result of Tables 4 and 5 it was concluded that there was significant correlation between error terms in regression and sample models for all data sets. So Heckman selection model was performed after these results.

In Table 6, selection model and regression model results can be seen. Beta coefficient values obtained in Heckman sample selection model formed for all three data set and confidence intervals values of these coefficients with *p* values used in testing of coefficients are given. When results are assessed, in the selection model, age was a significant factor for there data sets. According to this model for all three data sets, there was no significant relationship between Gleason score, smoking status and family history (for each p>0.05). But age,

TABLE 6: Results of Heckman sample selection model according to censored proportions.

			Coefficient	Standard Error	z	P>|z|	[95% Confidence Interval]	
Proportion of Benign tumor=0.30		Constant	-8.232	0.4718	-17.45	<0.001	-9.156	-7.307
	Gleason	Age	0.102	0.0047	21.70	<0.001	0.092	0.111
	(Regression	PSA	0.046	0.0029	15.87	<0.001	0.040	0.052
	Model)	Smoking	0.066	0.0425	1.56	0.119	-0.017	0.149
		Family History	0.024	0.0424	0.59	0.558	-0.058	0.108
		Weight	0.011	0.0024	4.43	<0.001	0.006	0.016
		Constant	-13.193	0.5382	-24.51	<0.001	-14.248	-12.138
	Selection	Age	0.173	0.0067	25.67	<0.001	0.160	0.186
	(Selection	Smoking	-0.074	0.0741	-1.01	0.313	-0.220	0.070
	Model)	Family History	0.058	0.0738	0.80	0.424	-0.085	0.203
Proportion of Benign tumor=0.50		Constant	-8.529	0.5391	-15.82	<0.001	-9.586	-7.472
	Gleason	Age	0.096	0.0050	19.26	<0.001	0.0866	0.106
	(Regression	PSA	0.042	0.0032	13.13	<0.001	0.035	0.048
	Model)	Smoking	0.090	0.051	1.78	0.075	-0.009	0.189
		Family History	0.034	0.051	0.67	0.500	-0.065	0.133
		Weight	0.017	0.003	5.98	<0.001	0.011	0.022
		Constant	-6.228	0.355	-17.53	<0.001	-6.924	-5.531
	Selection	Age	0.076	0.004	17.80	<0.001	0.067	0.084
	(Selection	Smoking	0.002	0.059	0.03	0.972	-0.115	0.119
	Model)	Family History	0.079	0.059	1.33	0.183	-0.037	0.196
Proportion of Benign tumor=0.70		Constant	-3.569	0.659	-5.41	<0.001	-4.862	-2.277
		Age	0.051	0.005	10.51	<0.001	0.041	0.060
	Gleason	PSA	0.022	0.003	6.58	<0.001	0.016	0.029
	(Regression	Smoking	0.008	0.055	0.15	0.882	-0.099	0.115
	Model)	Family History	-0.005	0.054	-0.08	0.934	-0.112	0.103
		Weight	0.019	0.004	4.49	<0.001	0.011	0.028
		Constant	-4.060	0.333	-12.17	<0.001	-4.714	-3.406
	Selection	Age	0.043	0.004	10.75	<0.001	0.035	0.051
	(Selection	Smoking	0.022	0.060	0.37	0.710	-0.095	0.140
	Model)	Family History	0.026	0.060	0.44	0.658	-0.091	0.144

psa and weight variables were significant risk factors in terms of Gleason score (for each p<0.001). For all three data sets in Heckman sample selection models (adding the mills ratio to sample selection model), positive beta coefficients have been obtained for age, psa and weight. This means that if these variables increase Gleason score also significantly increases. When beta coefficients of age, psa and weight variables are examined, for all three models the highest value has been observed to refer to age variable. For all three models standard values of beta coefficients have been observed to be similar. In summary, while smoking and family history does not have an effect in determining Gleason score, effects of psa, age and weight have been concluded to be statistically significant (Table 6).

DISCUSSION AND CONCLUSION

Heckman models couldn't find an extensive usage area in health field researches. In this study, the theoric explanations of Heckman sample selection model was given and an application was performed to give an example of the usage of the model in health studies.

When health literature is scanned this model has been observed to be used in a limited number of studies which have all been performed by foreign researchers. In PubMed, 16 studies in total have been found using this model, and 4 of these have been used for estimating HIV prevalence.

Galimard et al. (2016) has used Heckman's model in a randomise controlled clinical trial on seasonal influenza patients as an imputation method.[9] McGovern et al. (2015) and Clark and Houle (2014) have used Heckman's model to estimate HIV prevalence and they have stated that this model gives more consistent estimations for HIV prevalence.[5,14] DeMaris (2014) in his study, to consider also the effect of unmeasured confounding, has examined the association between being married and subjective well-being with Heckman's model. When model results are assessed, effect of unmeasured confoundings has been determined stronger and error term of the model has been indicated lower.[6] Ards et al. (1998) in order to examine the effects of sample selection bias on racial differences in child abuse reporting has used Heckman model and has concluded that sample selection effects estimations.[4]

So by looking these limited studies, we want to improve the knowledge of these models and to raise awareness about the importance of these models in health studies. It shall be kept in mind that the experimental designs where unmeasured confounding and selection bias are frequently seen are survey type researches or cross-sectional observational or clinical trials.

Consequently, in multifactorial health researches where important health decisions are taken, using the most appropriate statistical model will reduce model error and selection bias in obtained results.

Conflict of Interest

Authors declared no conflict of interest or financial support.

Authorship Contributions

All authors contributed equally to this study.

REFERENCES

1. Amemiya T. Tobit models: a survey. J Econom 1984;24(1-2):3-61.
2. Tobin J. Estimation of relationships for limited dependent variables. Econometrica 1995; 26(1):24-36.
3. Amemiya T. Regression analysis when the dependent variable is truncated normal. Econometrica 1997;41(6):997-1016.
4. Ards S, Chung C, Myers SL Jr. The effects of sample selection bias on racial differences in child abuse reporting. Child Abuse Negl 1998;22(2):103-15.
5. Clark SJ, Houle B. Validation, replication, and sensitivity testing of Heckman-type selection models to adjust estimates of HIV prevalence. PLoS One 2014;9(11):1-9.
6. DeMaris A. Combating unmeasured confounding in cross-sectional studies: evaluating instrumental-variable and Heckman selection models. Psychol Methods 2014;19(3):380-97.
7. Heckman J. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 1976;5(1):475-92.
8. Carson RT, Yixiao S. The tobit model with a non-zero threshold. Econom J 2007;10(3):1-15.
9. Galimard JE, Chevret S, Protopopescu C, Resche-Rigon M. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. Stat Med 2016;35(17): 2907-20.
10. Cragg J. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 1971; 39(5):829-44.
11. Chiburis R, Lokshin M. Maximum likelihood and two-step estimation of an ordered-probit selection model. Stata J 2007;7(1):167-82.
12. Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ. Cancer statistics, 2007. CA Cancer J Clin 2007;57(1):43-66.
13. Quinn M, Babb P. Patterns and trends in prostate cancer incidence, survival, prevalence and mortality. Part I: international comparisons. BJU Int 2002;90(2):162-73.
14. McGovern ME, Bärnighausen T, Salomon JA, Canning D. Using interviewer random effects to remove selection bias from HIV prevalence estimates. BMC Med Res Methodol 2015; 15(8):2-11.