Benefits of Threshold Regression: A Case-study Comparison with Cox Proportional Hazards Regression

Mei-Ling Ting Lee[1] and G. A. Whitmore[2] Bernard Rosner[3]

[1] University of Maryland, College Park, USA MLTLEE@UMD.EDU
[2] McGill University, Montreal, Canada name@email.address
[3] Harvard Medical School, Boston, MA, USA name@email.address

Abstract: Cox proportional hazards (PH) regression is a well-known model for analyzing survival data and its strengths are widely recognized. *Threshold regression* (TR) is a relatively new methodology but one that is receiving greater attention and being used successfully by researchers in different fields, including biopharmaceutical statistics. In threshold regression, event times are modeled by a stochastic process reaching a boundary threshold. The TR model does not require the proportional hazards assumption. It also can provide more insights into data than the Cox model, even where the PH assumption holds. Thus, threshold regression deserves consideration by investigators and their analysts as a serious alternative to Cox regression. In this article, we demonstrate the benefits of the TR model using a large cohort data set drawn from the Nurses' Health Study (NHS). The TR results for the NHS data set show the anticipated link between lung cancer and smoking for women. The TR model allows this link to be understood with substantial insight and clarity and with a refined attribution of disease progression to particular influences. We compare TR results with those obtained from Cox proportional hazards regression. The adequacies of the TR and Cox models in fitting the data set are examined using a new analytical approach. We also present STATA programs to compare the models.
Keywords: Endpoint, first hitting time, lifetime, maximum likelihood, lung cancer, smoking, stochastic process, survival analysis, threshold regression, time-to-event, Wiener diffusion process.
Acknowledgements: This research is supported in part by National Institutes of Health grant HL40619 (Rosner), OH008649 (Lee) and by a research grant from the Natural Sciences and Engineering Research Council of Canada (Whitmore).

1 Introduction

Cox proportional hazards (PH) regression is a well-known model for analyzing survival data and its strengths are widely recognized. *Threshold regression* (TR) is a relatively new methodology but one that is receiving greater attention and being used successfully by researchers in different fields, including biopharmaceutical statistics. In threshold

regression, event times are modeled by a stochastic process reaching a boundary threshold. The TR model does not require the proportional hazards assumption. It also can provide more insights into data than the Cox model, even where the PH assumption holds. Thus, threshold regression deserves consideration by investigators and their analysts as a serious alternative to Cox regression. Lee, Chang, Whitmore (2008) used a threshold regression mixture model for assessing treatment efficacy in a multiple myeloma clinical trial. They did not compare the benefits of TR with those of Cox PH regression.

In this article, we compare the TR and Cox models and demonstrate the benefits of the TR model using a large cohort data set drawn from the Nurses' Health Study (NHS). The TR results for the NHS data set show the anticipated link between lung cancer and smoking for women. The TR model allows this link to be understood with substantial insight and clarity and with a refined attribution of disease progression to particular influences. Specifically, we compare TR results with those obtained from Cox proportional hazards regression. The adequacies of the TR and Cox models in fitting the data set are examined using a new analytical approach. We also present STATA programs to compare the models.

2 First-hitting Time (FHT) and Threshold Regression (TR) Model

Threshold regression (TR) refers to a statistical model for time-to-event data in which the time to the event is defined as the first hitting time of an absorbing boundary by an underlying stochastic process. In our application of the TR model, the health status of each subject with respect to lung cancer follows a latent Wiener diffusion process $\{X(t)\}$ where t denotes time measured from the baseline of an observation interval. The initial health status of the subject at baseline is $X(0) = x_0 > 0$, which is a parameter to be estimated. The mean rate of change of health status over the interval is denoted by μ. Lung cancer occurs when the health status process first decreases to the zero-level, which is taken as an absorbing boundary or *threshold* for the process. The time of this first encounter, denoted by S, is called the *first hitting time* (FHT). If $\mu > 0$ then the lung cancer endpoint is not assured because the process would tend to drift away from the threshold. Other causes of death are competing with lung cancer and, hence, death from another cause will produce a right censored observation. In threshold regression, statistical techniques are used to estimate the effects of covariates on the parameters of the FHT model. See Aalen and Gjessing [2,3, 4] and Lee and Whitmore [4] for a review of FHT models and threshold regression.

2.1 The Nurses Health Study

We consider a large cohort data set drawn from the Nurses' Health Study (NHS) to compare the benefits of the TR model to the Cox model. The Nurses' Health Study (NHS) was established in 1976 when a cohort of 121,700 female registered nurses, aged 30 to 55 years, returned a mailed questionnaire reporting on disease history, personal characteristics and behaviors, and then updated the information by completing follow-up questionnaires on a biannual basis. The study was designed to allow prospective examination of the influences of lifestyle on the occurrence of disease, especially heart

disease and cancers. Every two years in follow-up questionnaires they have updated and extended these data. In this article, data from the NHS for the period 1986-2000 are considered. The data set consists of observation sequences for 115,768 women which represent 1,577,382 person-years at risk. The endpoint of interest here is a diagnosis of primary lung cancer as confirmed from medical records or death certificates. This endpoint was experienced by 1206 of the women by the year 2000. For more details about this study analyzed by the Cox PH model, see Bain, Feskanich *et al* [1]. We examine the link between lung cancer and smoking using the TR model and discuss the benefits of the TR model.

Our model assumes that the logarithm of initial health status $\ln(x_0)$ is a linear regression function of two covariates, namely, cumulative smoking at baseline *pkyrs0* (in pack years) and baseline age *age0* (in years). These covariates are selected for the initial health status on the assumption that initial health can only depend on conditions that prevail at baseline. The parameter μ describes the mean rate of change in health status with time. We assume that μ is a linear regression of the same covariates, *pkyrs0* and *age0*. In addition to these two covariates, we include a covariate that is an affine quadratic term for cumulative smoking, denoted by *pkyrs_sq*. The affine adjustment involves subtracting a constant from *pkyrs0* before squaring it. This adjustment reduces collinearity between the linear and quadratic terms. We have chosen the constant to be 28 pack years because this value reduces the correlation between *pkyrs0* and *pkyrs_sq* to almost zero. The quadratic term assesses the presence, if any, of a curvature effect in parameter μ for cumulative smoking. A further covariate, denoted by *dpkyrs*, is also included in the regression function for μ. This covariate represents the average annual rate of additional smoking by the subject between baseline and the endpoint (in pack years). The expectation is that this covariate will capture the influence on the rate of change in health status of continued smoking. Table 1 shows summary statistics for the failure indicator variable and the covariates.

Variable		Statistics			
Name	Units	Mean	Std. Dev.	Min.	Max.
Response					
fail	indicator (0,1)	0.0104174		0	1
Covariates					
age0	years	52.343	7.217	39	72
pkyrs0	pack years	13.072	18.456	0	122
pkyrs_sq	pk. yrs. squared	563.471	447.294	0	8836
dpkyrs	pack years	0.138	0.346	0	4

Table 1. Summary statistics for the failure variable and covariates used in the fitted threshold regression model. Covariate $pkyrs\_sq = (pkyrs0 - 28)^2$. All variables have 115,768 readings.

2.2 Model Estimation

The parameters of the TR model are estimated by maximum likelihood. Once the parameters are estimated, the values of the probability density function (p.d.f.) and cumulative distribution function (c.d.f.) for the interval from baseline to endpoint for

each subject can be estimated. It must be emphasized that the p.d.f. and c.d.f. are specific to lung cancer and take no account of mortality from other causes.

As already described, we let the latent health status process be a Wiener diffusion process $\{X(t)\}$. The FHT distribution in this setup is an inverse Gaussian distribution [5] and thus threshold regression corresponds to censored inverse Gaussian regression. The inverse Gaussian distribution depends on the mean and variance parameters of the underlying Wiener process (μ and σ^2) and the initial health status level (x_0). We let $f(t|\mu, \sigma^2, x_0)$ and $F(t|\mu, \sigma^2, x_0)$ denote the p.d.f. and c.d.f. of S, the time to a diagnosis of lung cancer. These functions have simple computational forms. For the case where the process begins at $x_0 > 0$ and the boundary is the zero level, the p.d.f. for S is given by

$$f(t|\mu, \sigma^2, x_0) = \frac{x_0}{\sqrt{2\pi\sigma^2 t^3}} \exp\left[-\frac{(x_0 + \mu t)^2}{2\sigma^2 t}\right], \quad -\infty < \mu < \infty,\ \sigma^2 > 0. \tag{1}$$

The c.d.f. corresponding to (1) is

$$F(t|\mu, \sigma^2, x_0) = \Phi\left[-\frac{(\mu t + x_0)}{\sqrt{\sigma^2 t}}\right] + \exp(-2x_0\mu/\sigma^2)\Phi\left[\frac{\mu t - x_0}{\sqrt{\sigma^2 t}}\right], \tag{2}$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution.

The event of experiencing a diagnosis of primary lung cancer in this study is equivalent to the process eventually experiences an FHT, i.e., the event $\{S < \infty\}$. We shall refer to the probability of this event, denoted by $P(S < \infty)$, as the subject's *probability of lung malignancy*. If the mean rate of change μ is positive then there is a possibility that the Wiener process will not experience the FHT and, in this case, the p.d.f. and c.d.f. in (1) and (2) are *improper* density and cumulative distribution functions. By taking the limit of the c.d.f. in (2) as t goes to infinity, it can be shown that

$$P(S < \infty) = \exp(-2x_0\mu/\sigma^2) < 1$$

if $\mu > 0$ and is 1 otherwise. The case where $\mu > 0$ features prominently in this study because we are considering only a single cause of death and eventual death from lung cancer is not a certainty.

For reference purposes, Appendix A.1 shows illustrative plots of the survival functions, probability density functions, and hazard functions for three improper inverse Gaussian distributions having $(x_0 = 1, \mu = .1)$, $(x_0 = 2, \mu = .2)$, and $(x_0 = 4, \mu = .4)$. These three distributions share the same conditional mean survival time $E(S|S < \infty) = 10$ but have different conditional variances. Their probabilities of lung malignancy $P(S < \infty)$ are 0.819, 0.449 and .041, respectively.

2.3 Regression Link Functions and Sample Log-likelihood Function

Because the health status process is latent here, it can be given an arbitrary measurement unit. Thus, in general, one parameter may be fixed and we choose to set the variance parameter σ^2 to unity. We link parameters μ and x_0 to baseline covariates that are represented by row vector $\boldsymbol{z} = (1, z_1, \ldots, z_k)$. The leading 1 in $\boldsymbol{z}$ allows for a constant term in the regression relationship. An identity link function of form

$$\mu = \boldsymbol{z}\boldsymbol{\beta} = \beta_0 + \beta_1 z_1 + \ldots + \beta_k z_k$$

is our choice for the mean parameter μ. A logarithmic link function

$$\ln(x_0) = \boldsymbol{z}\boldsymbol{\gamma} = \gamma_0 + \gamma_1 z_1 + \ldots + \gamma_k z_k$$

is our choice for the initial health parameter x_0. Here $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_k)'$, where β_0 and γ_0 are regression constants.

We now set up the sample log-likelihood function for censored inverse Gaussian regression. We assume that censoring is uninformative. Also, we assume that any nurse who contracts primary lung cancer during the study period does so at the end of the last reporting interval. We denote μ and x_0 for the ith subject by $\mu^{(i)}$ and $x_0^{(i)}$. We let $t^{(i)}$ denote the survival time of the ith subject for whom *fail* equals 1 or the right censoring time of the ith subject for whom *fail* equals 0. Hence, a subject i for whom *fail* = 1 contributes probability density $f(t^{(i)}|\mu^{(i)}, x_0^{(i)})$ to the sample likelihood function, for $i = 1, \ldots, n_1$, where $n_1 = 1206$ here. For subject i for whom *fail* = 0, the survival probability $\overline{F}(t^{(i)}|\mu^{(i)}, x_0^{(i)}) = 1 - F(t^{(i)}|\mu^{(i)}, x_0^{(i)})$ is the contribution to the sample likelihood function, for $i = n_1 + 1, \ldots, n_1 + n_0$. The sum $n = n_1 + n_0$, which equals 115,768 here, is the total number of subjects. Note that the variance parameter has been set to 1 and, hence, is suppressed in the preceding notation. The sample log-likelihood function to be maximized therefore has the form:

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n_1} \ln f(t^{(i)}|\mu^{(i)}, x_0^{(i)}) + \sum_{i=n_1+1}^{n_1+n_0} \ln \overline{F}(t^{(i)}|\mu^{(i)}, x_0^{(i)}). \tag{3}$$

Numerical gradient methods can be used to find maximum likelihood estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and estimates of their asymptotic standard errors. We have used a numerical optimization routine in *Stata* for this purpose. The *Stata* computational routines for this study are set out in Appendix A.2 and are seen to be quite simple.

3 Threshold Regression (TR) Investigations of Lung Cancer

Basic regression output

Table 2 shows output for our chosen regression model. A parsimonious model has been chosen to avoid overfitting and to simplify the interpretation of effects. For parameter $\ln(x_0)$, representing the logarithm of the initial health level, it is seen that covariate *pkyrs0* is significant, with a P-value of 0.000. Its regression coefficient is negative, signifying that baseline health status (with respect to lung cancer) tends to be lower for subjects with a larger amount of cumulative smoking at baseline. In other words, heavier smokers tend to be closer to a diagnosis of primary lung cancer. The covariate *age0* has a negative regression coefficient but it is not significant with the conventional 0.05 rule (a P-value of 0.066). For the mean parameter μ, the regression coefficients of all covariates are significant with P-values of 0.000. The coefficients of *age0*, *pkyrs0*, and *dpkyrs* are all negative, indicating the adverse effects on lung cancer health status of baseline age, baseline cumulative smoking, and continued smoking after baseline. The regression coefficient of *pkyrs_sq* for cumulative smoking is positive.

The combined linear and curvature effects for cumulative smoking suggest that heavier smoking is increasingly harmful to health but that the rate of increase moderates slightly with the amount of smoking.

Parameter	Variable	Estimate	Std. Error	P-value
$\ln(x_0)$				
	age0	-.0033470	.0018225	0.066
	pkyrs0	-.0030204	.0004752	0.000
	constant	1.792918	.1008379	0.000
μ				
	age0	-.0106827	.0012021	0.000
	pkyrs0	-.0036346	.0003566	0.000
	pkyrs_sq	.0000508	.0000058	0.000
	dpkyrs	-.1457822	.0086916	0.000
	constant	1.146989	.0674920	0.000

Table 2. Threshold regression output for a model in which parameter $\ln(x_0)$ depends on baseline age *age0* and baseline cumulative smoking *pkyrs0*. Parameter μ depends on the same covariates as well as an affine quadratic term for cumulative smoking *pkyrs_sq* and a covariate *dpkyrs* which represents the average annual smoking rate of the subject between baseline and the endpoint.

Estimated Risks of Developing Primary Lung Cancer

The fitted threshold regression model provides estimates for the absolute risk or probability of developing primary lung cancer for women of different baseline ages and smoking habits. Women in the cohort have been followed for at most sixteen years so the fitted model can only provide reliable estimates within a forward time horizon of about 15 years. Figure 1 shows the absolute risk of developing primary lung cancer within the next five years (panel a) and the next ten years (panel b) at different baseline ages *age0* for three smoking profiles: (1) a nonsmoker, (2) a smoker who has smoked one pack each day since age 18 and who will continue to smoke at the same rate, and (3) a smoker who has smoked two packs each day since age 18 and who will continue to smoke at the same rate. The figure shows the small risk of developing primary lung cancer for nonsmokers and the much greater risks for smokers, with the risk escalating with heavier smoking and advancing years. For the smoker of two packs per day, for example, the risk for the next decade of life rises to about 20 percent for women who are over 60 years old at baseline. The probabilities in Figure 1 take no account of competing risks of death and therefore will be larger than recorded mortality rates for lung cancer. It is clear that a potential death from lung cancer will not be observed and recorded for women who happen to die of other causes before the specified time horizon.

Another comparison of risks is offered by Figure 2. The figure shows estimated lung cancer survival functions over a 20-year horizon for a 45-year old nurse for four smoking profiles: (1) a nonsmoker, (2) a smoker who has smoked one pack each day since age 20 and who will continue to smoke at the same rate, (3) a smoker who has smoked two packs each day since age 20 and who will continue to smoke at the same rate, and (4) a smoker who has smoked one pack each day since age 20 but quits smoking at age 45.

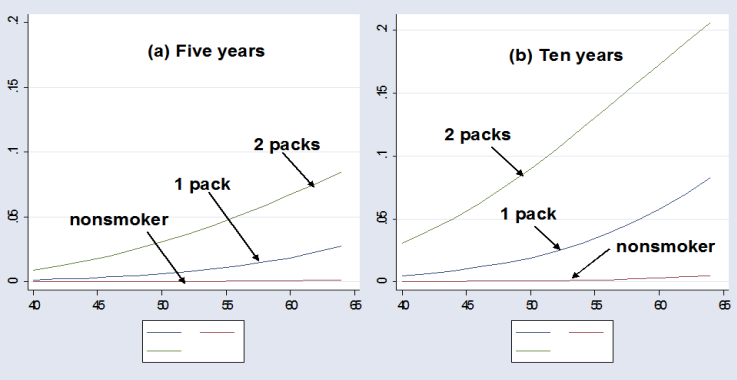

Fig. 1. The absolute risk (probability) of developing primary lung cancer within the next five years (panel a) and the next ten years (panel b) at different baseline ages *age0* for three smoking profiles: (1) a nonsmoker, (2) a smoker who has smoked one pack each day since age 18 and who will continue to smoke at the same rate, (3) a smoker who has smoked two packs each day since age 18 and who will continue to smoke at the same rate. The probabilities take no account of competing risks of death.

The survival curves take no account of competing risks of death. As expected, survival prospects are worst for the continuing smoker who has a two-pack per day habit. Her probability of developing primary lung cancer reaches close to 10 percent at the 20-year horizon when she would be 65 years of age.

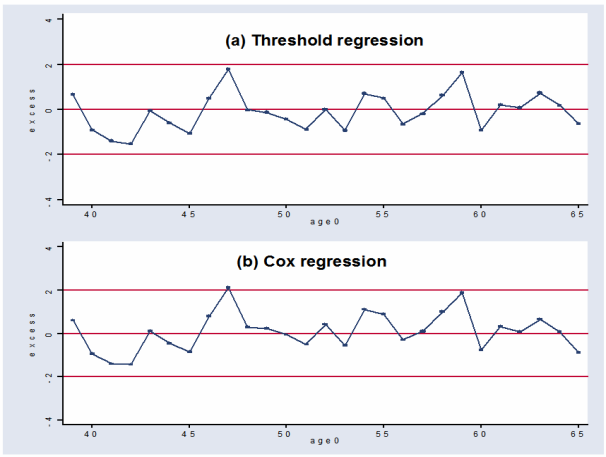

Fig. 2. Estimated lung cancer survival functions over a 20-year horizon for a 45-year old nurse for four smoking profiles: (1) a nonsmoker, (2) a smoker who has smoked one pack each day since age 20 and who will continue to smoke at the same rate, (3) a smoker who has smoked two packs each day since age 20 and who will continue to smoke at the same rate, and (4) a smoker who has smoked one pack each day since age 20 but quits smoking at age 45. The survival curves take no account of competing risks of death.

A final comparison of risks is offered by hazard functions. Figure 3 presents plots of hazard functions for three smoking profiles: (1) a nonsmoker, (2) a smoker who has smoked one pack each day since age 18 and who will continue to smoke at the same rate, (3) a smoker who has smoked two packs each day since age 18 and who will continue to smoke at the same rate. Panels (a) and (b) show these functions for nurses who are 40 and 60 years old at baseline, respectively. Observe that scales of the graphs are not comparable. The time horizon in each panel is 70 years of age and, thus, the hazard window in panel (a) is 30 years while that in panel (b) is 10 years. Also, the hazard levels in panel (b) are much larger than in panel (a) because the risk increases sharply with age.

A comparison of the ratios of the hazard functions for different smoking profiles over the age ranges presented in Figure 3 shows that the hazard functions are far from proportional. This observation is relevant for our comparison of TR and Cox proportional hazard regression that we address in a later section.

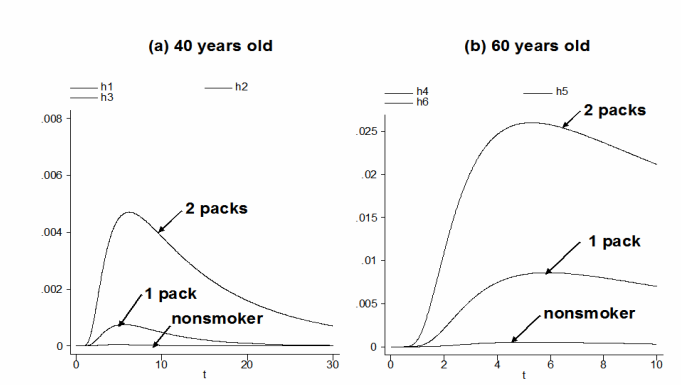

Fig. 3. Plots of hazard functions for three smoking profiles: (1) a nonsmoker, (2) a smoker who has smoked one pack each day since age 18 and who will continue to smoke at the same rate, (3) a smoker who has smoked two packs each day since age 18 and who will continue to smoke at the same rate. The functions in panel (a) span the 30-year period from a baseline age of 40 until age 70. Those in panel (b) span the 10-year period from a baseline age of 60 until age 70.

4 Comparisons of Results Obtained from the Cox PH Model

In this article, we compare TR results with those from the Cox proportional hazards regression model for the case of fixed covariates. The Cox model is the conventional one for this kind of application in time-to-event analysis - see, for example, Kalbfleisch and Prentice [6] and Cox and Oakes [7]. Comparisons of the TR model with the Cox regression for longitudinal data with time-dependent covariates will be discussed in a subsequent article.

$$h(t|\boldsymbol{\zeta z}) = h_0(t)\exp(\boldsymbol{\zeta z}) \tag{4}$$

Here $h(t|\boldsymbol{\zeta z})$ is the hazard function of a subject with covariate vector $\boldsymbol{z}$, $h_0(t) = h(t|0)$ is an arbitrary baseline hazard function, and $\boldsymbol{\zeta}$ is a vector of regression coefficients.

We have fitted the Cox model (4) to the data using the same covariates as for the TR model, namely, baseline age *age0*, baseline cumulative smoking *pkyrs0*, an affine quadratic term for cumulative smoking *pkyrs_sq*, and the average annual smoking rate of the subject between baseline and the endpoint *dpkyrs*. The regression results appear in Table 3. The signs of the regression coefficients for the covariates in Table 2 are the reverse of those in Table 3, which confirms that the effects for the covariates are in agreement with respect to the direction of effect. The signs of the regression coefficients in Table 3 show increasing hazard with increasing *age0*, *pkyrs0*, and continued smoking *dpkyrs*. The quadratic effect for cumulative smoking moderates the linear effect. Direct comparisons of the actual magnitudes of the coefficients are not meaningful, however, because they represent effects on parameters in completely different models.

Variable	Estimate	Std. Error	P-value
age0	.0967993	.005142	0.000
pkyrs0	.0434696	.0019047	0.000
pkyrs_sq	-.0005211	.000041	0.000
dpkyrs	1.0036	.0564842	0.000

Table 3. Cox proportional hazards regression results for the study using covariates: baseline age *age0*, baseline cumulative smoking *pkyrs0*, an affine quadratic term for cumulative smoking *pkyrs_sq*, and the average annual smoking rate of the subject between baseline and the endpoint *dpkyrs*.

4.1 Checking Model Fit

As a check on the TR model fitted in Table 2, we have examined the difference between actual and fitted lung cancer outcomes at different baseline ages *age0*. To compute the differences, we consider all subjects with a given baseline age a, where a ranges over 40 to 65 years. If $Y_a^{(j)}$ is an indicator variable for development of primary lung cancer for subject i of that age and $P_a^{(i)}$ is the true survival probability for the observation interval then the difference $Y_a^{(i)} - (1 - P_a^{(i)})$ has expected value 0 and variance $P_a^{(i)}(1 - P_a^{(i)})$. Our model is presumed to estimate the survival probability $P_a^{(i)}$ without bias. To check this claim, we have summed the differences $Y_a^{(i)} - (1 - \hat{P}_a^{(i)})$ at each distinct year of baseline age a, where $\hat{P}_a^{(i)}$ denotes the estimate of $P_a^{(i)}$. The approximate variance of this sum for each age is calculated as the sum of the individual variances based on the assumption that the sum components are independent. Independence is a reasonable approximation as the estimation errors for the parameters impart little dependence to the $\hat{P}_a^{(i)}$. The ratio of the sum of differences to its standard deviation at each age should be (approximately) a random standard normal number if the chosen TR model is correct. These ratios are plotted in panel (a) of Figure 4, with the points connected by straight lines to assist visual interpretation. The plot shows a zigsaw pattern that appears random and unbiased with no outliers. For a comparison, panel (b) of Figure

4 shows the standardized excess of actual lung cancer cases over predicted lung cancer cases for the fitted Cox regression model in Table 3. The zigsaw patterns are almost indistinguishable, showing that both models provide equally good fits to actual lung cancer outcomes as a function of baseline age.

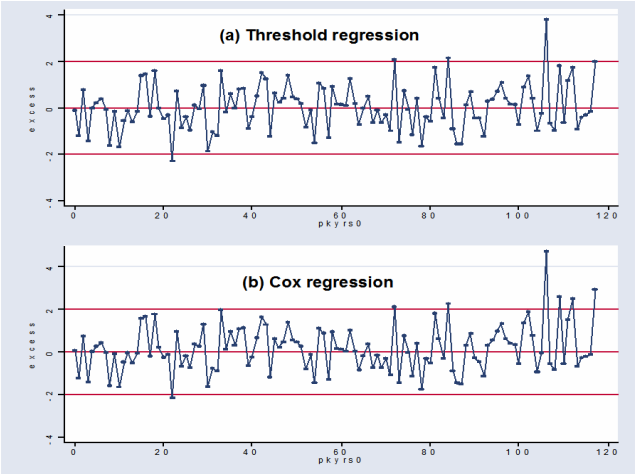

Fig. 4. Ratios representing the standardized excess of actual lung cancer cases over predicted lung cancer cases for each baseline year *age0* from 40 to 65 for (a) the fitted TR regression model in Table 2 and (b) the fitted Cox regression model in Table 3.

The check on fit was repeated with covariate *age0* replaced by the covariate *pkyrs0*. Figure 5 shows the resulting plots. Again we see similar fits, except for the region where *pkyrs0* exceeds about 100 pack years. In this upper region, the Cox model is biased as shown by the standardized excess being quite large for several points. Both graphs show one major outlier that happens to occur at the point where *pkyrs0* is 106.

A global test of the proportional hazards assumption was also performed. The result is a chi-square statistic of 1.59 for df = 4, giving a P-value of 0.811. The finding suggests the assumption is quite adequate for this application.

5 Benefits of Threshold Regression over Cox PH Regression

As noted already, the Cox proportional hazards (PH) regression model has been the model of choice for many studies involving time-to-event and survival data. We have seen that in spite of their different mathematical structures, threshold regression and Cox regression give qualitatively similar findings and similar fits to the data in this study. We quickly add that this similarity is not assured in other settings. Yet, in this setting, the clear similarity leads to the obvious question of why an alternative to Cox regression should be considered. We certainly recognize the strengths of Cox regression and, where its assumptions are valid, it should be used. On the other hand, there are benefits to threshold regression that should be considered by investigators and their analysts.

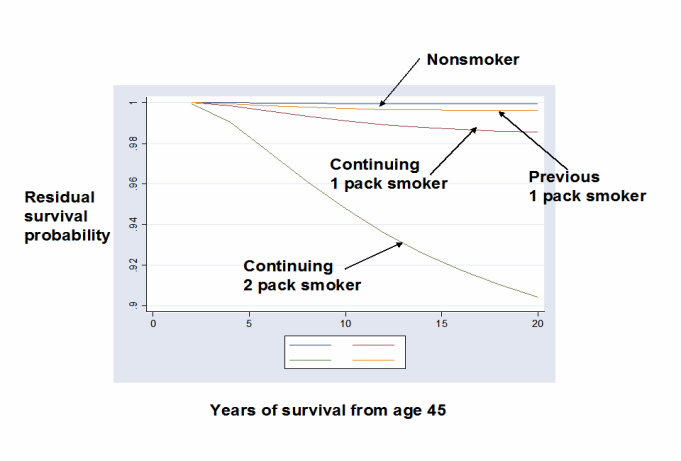

Fig. 5. Ratios representing the standardized excess of actual lung cancer cases over predicted lung cancer cases for each baseline cumulative pack years of smoking *pkyrs0* for (a) the fitted TR regression model in Table 2 and (b) the fitted Cox regression model in Table 3.

1. It can be shown that variants of the first hitting time model can be constructed that do have the PH property. Thus, adopting a threshold regression framework may enrich the interpretation of a Cox regression application. For example, setting Cox regression within a first hitting time context, if that is appropriate, can give a meaningful interpretation to the baseline hazard function.
2. Where a first hitting time model is appropriate and its survival functions do not have the PH property then threshold regression finds immediate application and the Cox model is disqualified. The inverse Gaussian survival distribution that is implicit in our TR regression application here does not possess the PH property. Yet, the Cox model and TR model do not differ sufficiently over the range of data to be statistically distinct. Women in this cohort have been monitored for only 16 years at most, which is not a long survival window.
3. The TR model is actually more parsimonious than the Cox regression model. The TR model is fully parametric and, in this application, has two set of coefficients for covariate effects, namely, those associated with the $\ln(x_0)$ and μ parameters. In contrast, Cox regression is a semi-parametric procedure because the baseline hazard function is arbitrary. The Cox model is rich in the parameters that define the baseline hazard function $h_0(t)$ and numerous degrees of freedom are absorbed in estimating that function, although this fact is not explicit in estimation routines based on the partial likelihood approach. Thus, the Cox model (4) involves estimating the regression coefficient vector $\boldsymbol{\zeta}$ as well as the baseline hazard function $h_0(t)$. Although the latter is often viewed as being of secondary interest, it deserves more attention than it receives. Criticism is sometimes leveled at investigators who use Cox regression without examining or attempting to understand the nature of its unspecified baseline hazard function.
4. As noted already, TR formulations force investigators to consider the actual causal mechanism of survival. Is a first hitting time involved? If so, what is the parent process? What is the nature of the absorbing boundary? What is the appropriate

regression structure for each parameter? Which covariates affect initial health status $\ln(x_0)$ and which influence the mean parameter μ that determines the course of disease progression after baseline. What are the relative magnitudes and directions of these influences? Threshold regression answers these important questions that are aimed at the scientific foundation of the analysis. In contrast, Cox regression provides only a regression structure for the log-hazard ratio without forcing investigators to dig deeper.

5. Most standard software packages contain a Cox regression routine. Therefore, some investigators may be reluctant to invest in programming the TR method. We point out, however, that TR requires little programming. For example, in *Stata* software, the programs used for this study involve only the following core program lines for TR and Cox regression.

   1. Threshold regression for the inverse Gaussian model

   ```
   ml model lf lung2008
   (lnx0: age0 pkyrs0)
   (m: age0 pkyrs0 pkyrs_sq dpkyrs)
   ```

   2. Cox regression with fixed covariates

   ```
   stset t, failure(f)
   stcox age0 pkyrs0 pkyrs_sq dpkyrs
   ```

   Observe that the Cox subroutine is built into *Stata* while TR regression requires a subroutine (called *lung2008* here). The subroutine computes the sample log-likelihood function in (3). The latter is not complicated in *Stata*, as it involves only specifying formulas for the normal density and distribution functions as shown in Appendix A.2, where the full TR program routines are presented.

References

1. Bain C, Feskanich D, Speizer FE, Thun M, Hertzmark E, Rosner BA, Colditz GA (2004). Lung cancer rates in men and women with comparable histories of smoking, *Journal of the National Cancer Institute*, **96 (11)**, June 2.
2. Aalen OO, Gjessing HK (2001). Understanding the shape of the hazard rate: a process point of view, *Statistical Science*, **16**, 1-22.
3. Aalen OO, Gjessing HK (2004). Survival models based on the Ornstein-Uhlenbeck process, *Lifetime Data Analysis*, **10**, 407-423.
4. Aalen OO, Borgan A, Gjessing HK (2008). Survival and Event History Analysis: A Process Point of View. New York: Springer.
5. Lee M-LT, Whitmore GA (2006). Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary, *Statistical Science*, **21**, 501-513.
6. Lee M-LT, Chang M, Whitmore GA (2008). A Threshold Regression Mixture Model for Assessing Treatment Efficacy in a Multiple Myeloma Clinical Trial, *Journal of Biopharmaceutical Statistics*, (In press).
7. Chhikara RS, Folks JL (1989). *The Inverse Gaussian Distribution: Theory, Methodology and Applications*, New York: Marcel Dekker.

8. Kalbfleisch JD, Prentice RL (1980). *The Statistical Analysis of Failure Time Data*, Wiley.
9. Cox DR, Oakes D (1984). *Analysis of Survival Data*, Chapman and Hall.

Appendix

A.1 Illustrative Plots of Improper Inverse Gaussian Distributions

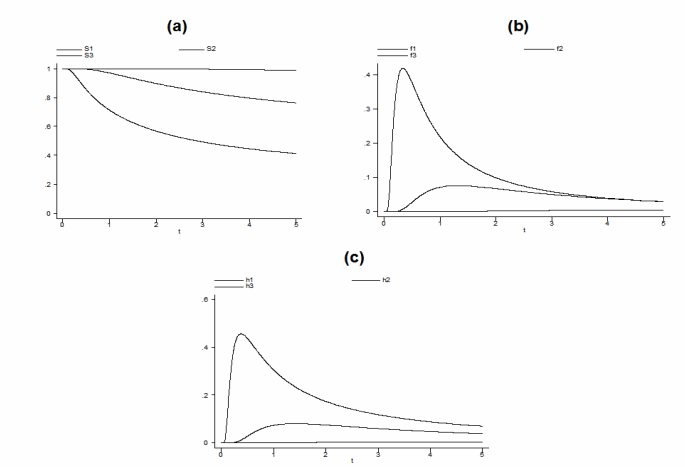

Fig. 6. Plots of the (a) survival functions, (b) probability density functions, and (c) hazard functions for improper inverse Gaussian distributions having $(x_0 = 1, \mu = .1)$, $(x_0 = 2, \mu = .2)$, and $(x_0 = 4, \mu = .4)$. These three distributions share the same conditional mean survival time $E(S|S < \infty) = 10$ but have different conditional variances. Their probabilities of malignancy $P(S < \infty)$ are 0.819, 0.449 and .041, respectively.

A.2 Stata Computer Routines

We use a numerical gradient method in *Stata* called *lf* to find the maximum likelihood parameter estimates and their estimated asymptotic standard errors given in Table 2. The main routine is listed below. The main routine calls a subroutine that we call *lung2008*. The main program also includes starting values for the numerical search. The subroutine *lung2008* is listed below the main program. The subroutine, at each iteration, computes the contribution of the current observation to the sample log-likelihood function. The symbols $f and $t denote the failure indicator and censoring or failure time of the current observation, respectively.

Main routine

```
ml model lf lung2008
(lnx0:age0 pkyrs0)
(m:age0 pkyrs0 pkyrs_sq dpkyrs);
ml init
```

```
0 0 2
0 0 0 0 1, copy;
```

Subroutine

```
program define lung2008
args lnf lnx0 m
tempvar x0
quietly gen double ‘x0’=exp(‘lnx0’)
quietly replace ‘lnf’= /*
*/ $f*(‘lnx0’-.5*(ln(2*_pi*($t^ 3))+(‘x0’+‘m’*$t)^ 2/$t)) /*
*/ +(1-$f)*ln(norm((‘m’*$t+‘x0’)/sqrt($t))/*
*/ -exp(-2*‘x0’*‘m’)*norm((‘m’*$t-‘x0’)/sqrt($t)))
end
```