Smoking Hazards: Evidence from the 2022 US National Youth Tobacco Survey

Econometric Methods for Empirical Economics Project Report

Anirudh Ravishankar

January 2024

Abstract

I study smoking hazards using data from the recent 2022 US National Youth Tobacco Survey (NYTS). I look at the age profiles by estimating Kaplan-Meier survival functions, and implement a split population duration model to study the determinants of starting smoking. I supplement my analysis by comparing the estimated results to those of the Cox proportional hazards model. I find that frequent social media use has a positive effect on the risk probability of smoking and the age at which individuals start smoking. This result is robust to multiple testing.

1 Introduction

During a time when the prevalence of cigarette smoking in the US has steadily been decreasing (Chen, Yu, & Wang, 2017), especially amongst the youth, it is interesting to study how the determinants of starting smoking have evolved since earlier work on the subject. Although it is plausible that the rising popularity of a variety of smoking substitutes such as e-cigarettes, vapes, etc. have influenced this declining trend, a study of competing risks is beyond the scope of my inquiry. Instead, the main motivation for this study is that revisiting the fundamental question of the factors affecting cigarette smoking uptake using new data can reveal new insights and interesting comparisons to the literature.

An important and common goal behind studies of smoking are implications for policies designed to reduce smoking among the population. Prices are one factor that have long been hypothesised to play an important role in individuals' decisions to smoke. But empirical studies studying prices have found conflicting results. Forster and Jones (2001) study the effect of a proposed tax on cigarettes in the UK on individuals' number of years smoked, and find a significant effect of higher taxes on reducing the years that men and women smoke. Earlier, Douglas and Hariharan (1994) studied the determinants of the hazard of starting smoking for a sample of US adults including the price elasticity of cigarettes. They find no effect of prices, but find that non-economic variables such as gender, race, stress, etc. significantly explain the decision to start smoking.

Consequently, models of addiction have evolved to explain that price shocks are unlikely to reduce consumption due to adjustment costs, and rationalise results such as those of Forster and Jones (2001) as stemming from the use of aggregate price data (Suranovic, Goldfarb, & Leonard, 1999, 2001). For this reason, I do not assess the effect of price variation in this study. Furthermore, I use self-reported data from the US National Youth Tobacco Survey (NYTS) which finds that approximately only half of respondents state that their access to tobacco products is by buying them or having someone buy the product for them (Gentzke et al., 2022). This is likely due to the fact that the survey focuses on school kids, whose purchasing power could be highly confounded. Thus, the addition of prices as an explanatory variable may obscure any direct causality while doing the estimation.

Following the methodology in Douglas and Hariharan (1994), I implement a split population duration model (Schmidt & Witte, 1989) to examine the determinants of the probability of smoking and age of starting smoking respectively. I compare the results to those obtained from the Cox proportional hazards model, a semi-parametric approach restricted to determining only the timing of starting smoking. The main finding is that frequent social media use has a positive effect on the probability of risk and timing for an individual.

A well known limitation of using survey data for duration analysis is recall bias, a disadvantage which I believe may be curbed by the choice of data I have at hand. The mean age for starting smoking has been found to be before the age of 20 (Pérez et al., 2021), hence, recall bias may be attenuated to some extent by using data collected from individuals around this age.

This paper is structured as follows: Section 2 and Section 3 present the data used and the empirical model respectively. Section 4 presents the results obtained for the Cox model and split population model, and Section 5 concludes. My R script for conducting the analysis is presented in the Appendix.

2 Data

I obtain data from the 2022 NYTS, which serves as a comprehensive and nationally representative cross-sectional dataset.[1] It contains information on the behaviors and attitudes of school-age youth, primarily focusing on the use of tobacco products and extends to tobacco products containing some other addictive substances (Gentzke et al., 2022; Park-Lee et al., 2022). In addition to capturing information on tobacco use, it includes variables for demographic characteristics such as sex, ethnicity, sexual identity, and social determinant indicators such as self-reported mental health measures, wealth (number of cars in the family, whether the respondent has their own bedroom, vacation frequency, etc.) and school grades.

A few comments can be made *ex ante* with regard to the validity of using such a dataset. First, the data was collected in the post COVID-19 period, which may influence the degree of comparison that can be made using the results I obtain using it. Second, although nationally representative among school-going youth, the data does not reflect information about individuals outside this group. Most children in the age group 10-17 are enrolled in schools.[2] However, it is impossible to know how much results in the empirical literature are driven by the latter group.

For each tobacco product, participants were asked the following questions (paraphrased): (i) "Have you ever used this product?" and (ii) "How old were you when you first used this product?" To study the hazards of starting smoking, the following variables are considered from the dataset for each individual:

AGE, the individual's age in years at time of interview. This is an integer value in the range 9 to 19.

MALE, a dummy variable equal to one if the individual is male, and equal to zero if female.

ETHNICITY, a categorical variable indicating if the individual belongs to one of the following groups – Native Indian, Asian, Black, Hawaii/Pacific, White, Hispanic, or Mixed.

SOCIAL_MEDIA_USE, a categorical variable indicating the individual's current frequency of social media use – Never, Weekly, or Daily.

NO_OF_CARS, a categorical variable indicating the number of cars currently owned by the individual's family – 0, 1, or >= 2 (i.e., 2 or more).

OWN_BEDROOM, a dummy variable equal to one if the individual reports having their own bedroom, and equal to zero otherwise.

SCHOOL_GRADES, a categorical variable indicating the individual's current grades – High (A's or B's), or Low (C's or below).

CIGARETTE_EVER, a dummy variable equal to one if the individual reports ever having used a cigarette up until the time of interview, and equal to zero otherwise.

EXIT_AGE, the age in years at which the individual first used a cigarette (if they ever did so), or the individual's age at time of interview if they have never smoked. This variable is an integer in the range 8 to 19.

[1] https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/data/index.html

[2] https://www.statista.com/statistics/236087/us-school-enrollment-rates-by-age-group/

EXIT_AGEC, the age in years when the individual started smoking, conditional on starting. This variable is not in the dataset (although I do calculate it in the R script). I present it to help get a better descriptive picture of the data.

Table 1 presents summary statistics for the full sample and the analytic sample.[3]

Table 1: Summary statistics. [N (%) or Mean (min, max)]

Variable	*Full sample:* N = 22,690	*Analysis sample:* N = 2,269
AGE	14.61 (9.00, 19.00)	14.61 (9.00, 19.00)
MALE		
0	11,317 (50%)	1,137 (50%)
1	11,373 (50%)	1,132 (50%)
ETHNICITY		
Mixed	6,302 (28%)	619 (27%)
Native Indian	853 (3.8%)	89 (3.9%)
Asian	1,586 (7.0%)	135 (5.9%)
Black	2,415 (11%)	227 (10%)
Hawaii/Pacific	123 (0.5%)	11 (0.5%)
White	10,327 (46%)	1,073 (47%)
Hispanic	1,084 (4.8%)	115 (5.1%)
SOCIAL_MEDIA_USE		
Never	1,751 (7.7%)	160 (7.1%)
Weekly	2,058 (9.1%)	197 (8.7%)
Daily	18,881 (83%)	1,912 (84%)
NO_OF_CARS		
0	641 (2.8%)	57 (2.5%)
1	3,239 (14%)	296 (13%)
>= 2	18,810 (83%)	1,916 (84%)
OWN_BEDROOM		
0	3,899 (17%)	382 (17%)
1	18,791 (83%)	1,887 (83%)
SCHOOL_GRADES		
High	18,507 (82%)	1,854 (82%)
Low	4,183 (18%)	415 (18%)
CIGARETTE_EVER	1,888 (8.3%)	189 (8.3%)
EXIT_AGE	14.37 (8.00, 19.00)	14.34 (8.00, 19.00)
EXIT_AGEC	12.61 (8.00, 18.00)	12.26 (8.00, 18.00)

The raw data contains 28,291 observations, but after cleaning it I obtain 22,690 observations. Individuals with incomplete information on the required variables were removed, and the indicator variables from the data were collapsed into the respective categorical variables described previously.

3 Empirical Methodology

First, I look at the empirical distributions of the data by plotting the survival and hazard functions. The survival function $P(t)$, defined as the probability of 'surviving' (i.e., not experiencing the event of interest) beyond time t, is given by

$$P(t) = Pr(T > t) \tag{1}$$

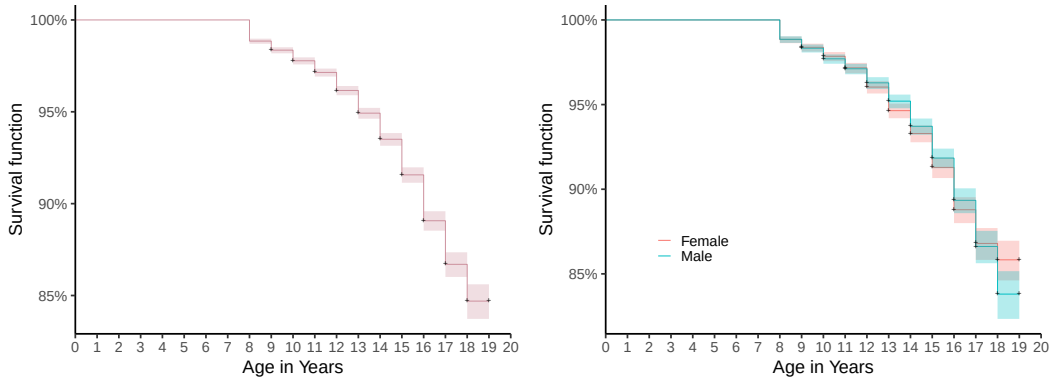
[3]See Section 3 on how I implement the main empirical model.

where T is the last period of observation, which is usually the time of surveying the individual. One way of estimating the survival function non-parametrically is using the Kaplan-Meier estimator (Kaplan & Meier, 1958), given by

$$\hat{P}(t) = \prod_{s<t} \big(1 - \hat{h}(s)\big) \tag{2}$$

where $h(s)$ is the hazard function at a time $s < t$. The hazard function, or the rate of hazard, is the primary focus of interest in duration analyses. It is defined as the instantaneous probability of experiencing the event, which in this case is the age of starting cigarette smoking. From Equation 2, it can be seen that the respective survival and hazard functions are transformations of each other, and affirm the same information, although the hazard is more easy to interpret since it is a probability.

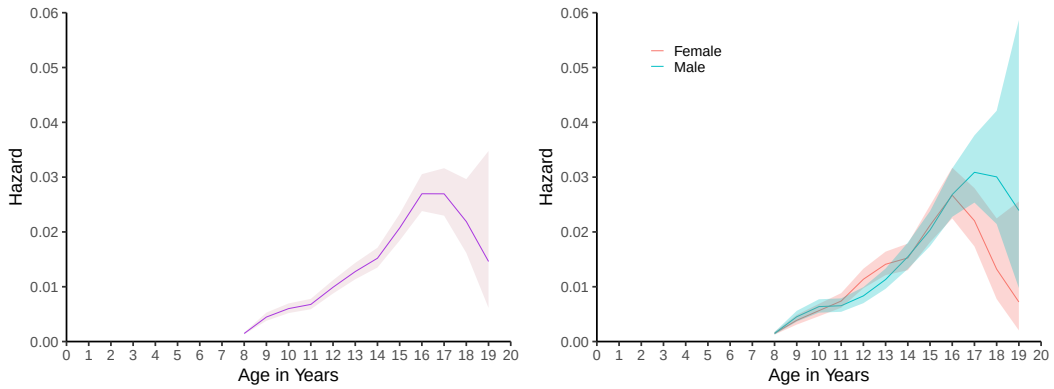
Figure 1a and Figure 1b show the Kaplan-Meier survival curves for the full sample and by sex respectively. Since the data on the age at which smoking began only contains integer values, the curves appear to be highly discretised.



(a) Global. (b) By sex.

Figure 1: Survival curves.

Figure 2a and Figure 2b present the empirical hazards for the full sample and by sex respectively.



(a) Global. (b) By sex.

Figure 2: Empirical hazard curves.

The distribution of the hazard influences the modelling choice. Evidently, the hazard in this case is non-monotonic, as also seen by Douglas and Hariharan (1994)

i.e., the curve appears to increase at first and decreases after reaching a peak. First, without any distributional assumptions on the hazard, I begin by estimating the semi-parametric Cox proportional hazards model (Cox, 1972). The model is typically represented as

$$h(t; x_1, x_2, \ \ldots) = h_0(t) \times e^{\beta_1 x_1 \ + \ \beta_2 x_2 \ + \ \ldots \ + \ \beta_k x_k} \tag{3}$$

where x_1, x_2, ..., x_k are explanatory variables, β_1, β_2, ..., β_k are parameters, $h(t)$ and $h_0(t)$ are the hazard rate (defined earlier) and baseline hazard (i.e., hazard for a null model) respectively. e^{β_1}, e^{β_2}, ..., e^{β_k} are called hazard ratios. A hazard ratio above one indicates a positive relation with the event probability, and hence a negative association with the survival duration. Essentially, the model estimates the effect on the age at which the event occurs i.e., whether it occurs sooner or later in life (Bartels, 2003).

The validity of the Cox model relies on the assumption of proportional hazards, which states that the hazards for the levels of a covariate must be proportional over time. Log-rank test results for the violation of this assumption are given in Table 2. For example, sex does not violate the assumption as seen in Figure 2b. If any variable violates the proportional hazards assumption, they are included as a stratified covariate in the estimation.[4] For the sake of integrity, I report results for both the stratified and non-stratified model specifications. Furthermore, to correct for false positives in multiple comparisons, I employ the false discovery rate (FDR) method (Benjamini & Hochberg, 1995).

Table 2: Violation of proportional hazards.

Group	Log-rank p	PH violated
Sex	0.61	No
Ethnicity	0.00	Yes
Social media use	0.11	No
No. of cars	0.00	Yes
Own bedroom	0.00	Yes
School grades	0.00	Yes

Next, I estimate a split population duration model on the NYTS data. The motivation behind this model is the violation of the assumption in standard duration models, such as the Cox model, that every individual will experience the event of interest (denoted as 'failure') eventually (Schmidt & Witte, 1989). From Table 1, it is evident that only 8.3% of the entire sample ever smoke, corresponding to 1,888 individuals out of 22,690. Therefore, it is imperative to account for the probability that a given individual will even begin smoking in their lifetime.

Adapting notation from Douglas and Hariharan (1994), the likelihood function for individual i in the model is given by

$$\begin{aligned} L_i(\theta, \beta; T_i, z_i, x_i) = \ & \delta_i G(z_i, \theta) f(T_i; x_i, \beta) \\ & + (1 - \delta_i)[(1 - G(z_i, \theta)) + G(z_i, \theta) S(t; x_i, \beta)] \end{aligned} \tag{4}$$

where z_i and x_i are covariates affecting the probability and duration of smoking respectively, T_i the time until smoking, β and θ parameters, δ_i an indicator equalling one if the individual has ever smoked up until time t, $G(\cdot)$ is the probability that

[4] See my R script.

the individual begins smoking, $f(\cdot)$ is the density (log-logistic in this case) at T_i, and $S(\cdot)$ is the survival function.

A log-logistic distribution of the hazard is assumed due to the fact that the observed empirical hazard increases first and then decreases. Since there is no data beyond age 19, it cannot be said for certain but the shape of the curve in Figure 2a appears to conform to the literature, reaching a peak before the age of 20. With this distributional assumption, the hazard function can thus be written as

$$h(t) = \frac{f(t)}{S(t)} = \frac{(\gamma/\alpha)(t/\alpha)^{\gamma-1}}{1+(t/\alpha)^{\gamma}} \tag{5}$$

where $f(\cdot)$ is the probability density function of the log-logistic distribution, α and γ are scale and shape parameters respectively.

I use the same covariates in z_i and x_i i.e., I study the effects of the same variables (outlined in the previous section) on both probability and duration of smoking. One may not be convinced of the validity of the variables *SOCIAL_MEDIA_USE*, *NO_OF_CARS*, *OWN_BEDROOM*, and *SCHOOL_GRADES* in explaining the the decision to smoke since they are collected at the time of interview and thus may not indicate the conditions in the life of an individual at the time they first smoked. For this reason, I estimate two specifications of the model: one with only *MALE* and *ETHNICITY*, and one with all covariates. For the latter, I make the (big) assumption that these variables were the same for the individual when they first smoked.

Of the explanatory variables, only *NO_OF_CARS* is an economic variable that proxies wealth. My justification for the inclusion of the rest are as follows: First, a substantial body of literature has documented empirical evidence indicating a positive association between frequent engagement with social media platforms and individuals' attitudes towards electronic cigarettes and related products (Alpert, Chen, & Adams, 2020; Vogel et al., 2021). Second, having one's own bedroom during teenage years can help develop a sense of lifestyle and independent identity (Bovill & Livingstone, 2013). I postulate that the presence of a bedroom is an enabler to indulge in smoking. Finally, for school grades, Park-Lee et al. (2022) in their descriptive analysis of the 2022 NYTS find an inverse relationship between prevalence of tobacco use and school grades.

Following Schmidt and Witte (1989) and with regard to computational limitations, I split the study sample in the following manner – a 10% random sample of the population is designated the 'analysis sample' (corresponding to 2,269 individuals, as shown in Table 1), while the rest is the 'test sample'. The model is estimated on the analysis sample, and the test sample is used for out-of-sample predictions to test the model's accuracy.

4 Results

To begin with, an interesting point of note is sex differences in the age profile of smoking. It can be observed from Figure 1b that the survival functions of males and females are similar, only differing at age 19 with comparatively more males taking up smoking. Figure 2b is more revealing, showing a clear disparity between males and females after age 16, with males subject to higher risk. However, the hazard curves have large confidence intervals at later ages due to smaller sample size, and

the test for violation of proportional hazards for sex is non-significant. Therefore, this observation is subject to caution.

Table 3 presents the results for the stratified and non-stratified Cox models respectively. The coefficients are non-exponentiated, and thus taking their exponent gives the hazard ratio (defined in previous section) for that covariate. It is interesting to note that the most valid result is that higher social media use has a positive effect, albeit by a small magnitude. This is consistent with the aforementioned literature on social media influence. In the non-stratified model (which does not account for violation of proportional hazards) interpretability is limited. Still, the coefficients for ethnicity and school grades conform to the survey finding that use of tobacco products was higher for non-Hispanic American Indian or Alaska Natives (grouped together as Native Indian in my data), and individuals with low school grades (Gentzke et al., 2022).

Table 3: Cox proportional hazards model.

	(1) *Stratified*	(2) *Non-stratified*
MALE		
1	−0.087* (0.047)	−0.087* (0.047)
ETHNICITY		
Native Indian		0.676***† (0.092)
Asian		−0.856***† (0.138)
Black		−0.614***† (0.098)
Hawaii/Pacific		−0.268 (0.319)
White		0.059 (0.055)
Hispanic		−0.345***† (0.128)
SOCIAL_MEDIA_USE		
Weekly	0.175 (0.127)	0.173 (0.126)
Daily	0.253**† (0.103)	0.255**† (0.102)
NO_OF_CARS		
1		−0.202 (0.123)
>= 2		−0.441***† (0.114)
OWN_BEDROOM		
1		−0.123** (0.061)
SCHOOL_GRADES		
Low		0.734***† (0.050)
Observations	22,690	22,690
Log Likelihood	−12,095.88	−17,854.17
LR Test	11.46*** (df = 3)	439.25*** (df = 13)
Score (Logrank) Test	11.01** (df = 3)	492.226*** (df = 13)
Note: Standard errors in parentheses.	† *FDR Robust.*	*p<0.1; **p<0.05; ***p<0.01

Table 4 presents the results of the split population model, for the two specifications described in the previous section. There are several comments to be made. First, it is evident that model specification changes the estimates. None of the coefficients for the "safe" model are significant for either duration of smoking or probability. On the other hand, the model with all covariates produces interesting values. Second, as found in the Cox model, for ethnicity, Native Indians tend to take up smoking earlier in life and Hispanics are at less risk. And again, high frequency social media use has a significant positive effect on the risk of smoking. But it also has a positive effect on the duration! This means that individuals that use social media a lot start smoking later in life. Furthermore, the result for low grades does not reflect that of the Cox model. Finally, having one's own bedroom does

not appear to have any effect as postulated.

Table 4: Split population model.

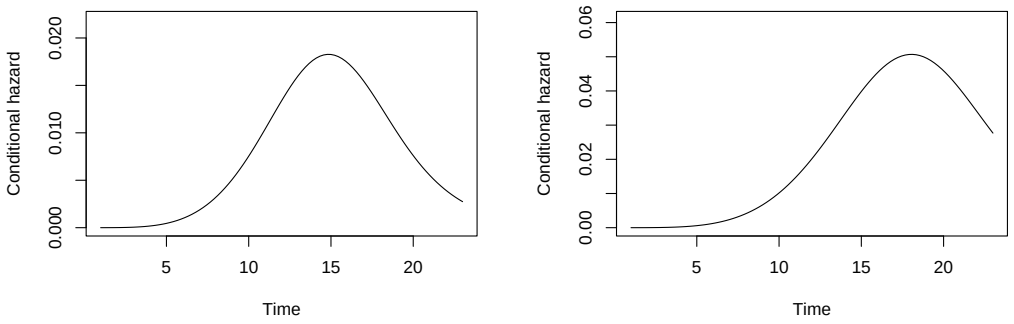
	Duration results:	
	(1)	(2)
Intercept	2.78*** (0.26)	1.87*** (0.26)
MALE		
1	0.14 (0.16)	0.11 (0.08)
ETHNICITY		
Native Indian	-0.27 (0.27)	-0.64***† (0.17)
Asian	-0.11 (0.55)	-0.33 (0.37)
Black	-0.44 (0.29)	0.27* (0.14)
Hawaii/Pacific	1.87 (23.33)	2.43 (323.41)
White	0.14 (0.26)	-0.12 (0.10)
Hispanic	-0.35 (0.45)	-0.60 (0.41)
SOCIAL_MEDIA_USE		
Weekly		0.56** (0.26)
Daily		1.08***† (0.24)
NO_OF_CARS		
1		0.36** (0.18)
>= 2		0.13 (0.14)
OWN_BEDROOM		
1		0.04 (0.10)
SCHOOL_GRADES		
Low		-0.08 (0.09)
	Risk probability results:	
	(1)	(2)
Intercept	-0.96 (1.20)	-3.69* (1.86)
MALE		
1	0.36 (0.89)	0.59 (0.78)
ETHNICITY		
Native Indian	0.79 (1.38)	-3.94* (1.98)
Asian	-1.82 (2.26)	-6.01** (2.71)
Black	-1.79 (1.31)	1.29 (2.16)
Hawaii/Pacific	0.00 (3.10)	-0.15 (1923.99)
White	0.49 (1.56)	-3.18** (1.55)
Hispanic	-2.22 (1.68)	-7.00**† (2.59)
SOCIAL_MEDIA_USE		
Weekly		3.56** (1.80)
Daily		7.01***† (2.32)
NO_OF_CARS		
1		1.35 (1.79)
>= 2		-0.79 (1.23)
OWN_BEDROOM		
1		0.29 (0.92)
SCHOOL_GRADES		
Low		2.02* (1.10)
log(α)	-1.66*** (0.16)	-1.64*** (0.12)
AIC	649.81	616.31

Note: Standard errors in parentheses. † *FDR Robust.* *p<0.1; **p<0.05; ***p<0.01

Some comparisons can be made to the results of Douglas and Hariharan (1994). Since they exclude individuals younger than age 25 (to see the effect of completed education), the study populations are not similar in the sense that the ages do not overlap at all. They mainly find that males smoke earlier in life and are at higher

risk. Also, being Black has a negative effect on the probability of smoking and a positive effect on starting later in life. I do not find any significant result for males, and find different results for ethnicity, which may have arisen from a shifting demographic over the years (their study is thirty years old). As far as interpretability goes, it would be ideal to use the entire dataset I have to estimate the model and obtain better coefficients. But unfortunately, I reiterate that computational reasons had to be considered.

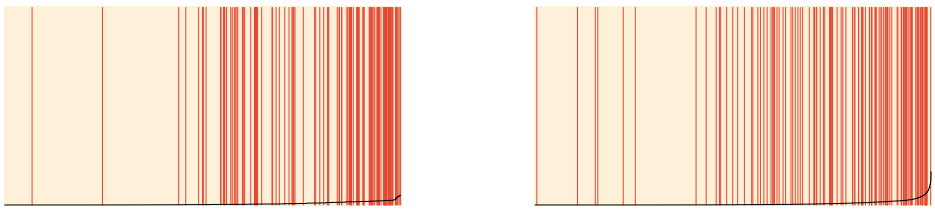
The silver lining lies in checking the model fit. According to the AIC, the specification with all covariates appears to be a better fit for the data. In addition, Figure 3a and Figure 3b show the estimated hazard functions from the model for both specifications. The shape of the hazard appears strikingly similar to the empirical hazard in Figure 2, leading me to believe that the distributional assumption was accurate. However, looking at the magnitudes, the models either underestimate or overestimate the true hazard. The model with all covariates fares better overall, appearing to peak at the same point as the empirical hazard.



(a) Restricted model. (b) Unrestricted model.

Figure 3: Log-log estimated hazards.

Next, model predictions are done on the test sample (described in the previous section) using separation plots (Greenhill, Ward, & Sacks, 2011). Figure 4a and Figure 4b plot the predictions for the assumed log-logistic distribution model. Events or failures are shown as red lines, while non-events are shown in light yellow. The arrangement of observations from left to right is determined by the predicted probabilities, with higher values towards the right.



(a) Restricted model. (b) Unrestricted model.

Figure 4: Model prediction.

The plots aid in assessing the ability of the model to accurately predict outcomes by evaluating its performance in assigning probabilities relative to the observed events. It is evident that both model specifications predict well on the test sample.

5 Conclusion

In this paper, I aimed to study the determinants of starting cigarette smoking among US school-age youth using a new dataset, the 2022 National Youth Tobacco Survey (NYTS). Mainly following the methodology in Douglas and Hariharan (1994), I use a split population duration model to study the hazard of smoking, and compare results with a Cox proportional hazards model. The only valid point of comparison between their study and mine is the significant effect of ethnicity variables. Other than that, the main finding is that frequent social media use emerges as a factor positively influencing both the probability and timing of smoking initiation. This finding aligns with existing literature linking social media use to positive attitudes towards tobacco products.

In comparing model specifications, I acknowledge variation in estimates, which underscores the importance of careful consideration in model selection. Furthermore, I take measures to improve the robustness of the estimates. In case of the Cox model, I test for the validity of the proportional hazards assumption. And in the split population model, I study the estimated hazard and check the strength of out-of-sample predictions. Finally, the estimates I interpret are robust to multiple testing using the false discovery rate (FDR) method.

References

Alpert, J. M., Chen, H., & Adams, K.-A. (2020). E-cigarettes and social media: attitudes and perceptions of young adults to social media messages. *Addiction research & theory*, *28*(5), 387–396.

Bartels, B. (2003). Advances in duration modeling: The split population duration model. *Lab Notes*, *3*, 4–7.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Bovill, M., & Livingstone, S. (2013). Bedroom culture and the privatization of media use. In *Children and their changing media environment* (pp. 179–200). Routledge.

Chen, X., Yu, B., & Wang, Y. (2017). Initiation of electronic cigarette use by age among youth in the us. *American journal of preventive medicine*, *53*(3), 396–399.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Douglas, S., & Hariharan, G. (1994). The hazard of starting smoking: estimates from a split population duration model. *Journal of health economics*, *13*(2), 213–230.

Forster, M., & Jones, A. M. (2001). The role of tobacco taxes in starting and quitting smoking: duration analysis of british data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *164*(3), 517–547.

Gentzke, A. S., Wang, T. W., Cornelius, M., Park-Lee, E., Ren, C., Sawdey, M. D., ... Homa, D. M. (2022). Tobacco product use and associated factors among middle and high school students—national youth tobacco survey, united states, 2021. *MMWR Surveillance Summaries*, *71*(5), 1.

Greenhill, B., Ward, M. D., & Sacks, A. (2011). The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science*, *55*(4), 991–1002.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481.

Park-Lee, E., Ren, C., Cooper, M., Cornelius, M., Jamal, A., & Cullen, K. A. (2022). Tobacco product use among middle and high school students—united states, 2022. *Morbidity and Mortality Weekly Report*, *71*(45), 1429–1435.

Pérez, A., San N'hpang, R., Callahan, E., Bluestein, M., Kuk, A. E., Chen, B., ... Harrell, M. B. (2021). Age at initiation of cigarette use in a nationally representative sample of us youth, 2013-2017. *JAMA network open*, *4*(2).

Schmidt, P., & Witte, A. D. (1989). Predicting criminal recidivism using 'split population' survival time models. *Journal of Econometrics*, *40*(1), 141–159.

Suranovic, S. M., Goldfarb, R. S., & Leonard, T. C. (1999). An economic theory of cigarette addiction. *Journal of health economics*, *18*(1), 1–29.

Suranovic, S. M., Goldfarb, R. S., & Leonard, T. C. (2001). Are rival theories of smoking underdetermined? *Journal of Economic Methodology*, *8*(2), 229–251.

Vogel, E. A., Ramo, D. E., Rubinstein, M. L., Delucchi, K. L., Darrow, S. M., Costello, C., & Prochaska, J. J. (2021). Effects of social media on adolescents' willingness and intention to use e-cigarettes: an experimental investigation. *Nicotine and Tobacco Research*, *23*(4), 694–701.

Appendix

Here I present my commented R script, which can also be found on my GitHub. First I load the relevant libraries, then import and clean the data.

```
# Econometric Methods for Empirical Economics Project
# Anirudh Ravishankar
# January, 2024

# Libraries ------------------------------------------------------

library(dplyr)
library(ggplot2)
library(readxl)
library(tidyr)
library(ggfortify) # Survival plots
library(bshazard) # Hazard plots

# Proportional hazards models
library(survival)
library(eha)

# Split population duration model
library(spduration)
library(separationplot)

# False discovery rate (FDR)
library(FDRestimation)

# Tables
library(gtsummary)
library(stargazer)
library(xtable)

# Data Import and Cleaning ---------------------------------------

# NYTS 2022 Data (takes about 70 seconds to load on my system)
#
# QN1: Age at interview
# QN2: Sex
#
# Ethnicity -
# QN4B, C, D, E: Hispanic
# QN5A: American Indian or Alaska Native
# QN5B: Asian
# QN5C: Black
# QN5D: Hawaii/Pacific
# QN5E: White
#
# Interesting covariates -
# QN133: How often you use social media?
# QN161: Wealth. Does family own cars?
# QN162: Wealth. Do you have your own bedroom?
# QN165: School grades currently.
#
# Cigarettes -
# QN35: Ever smoked?
# QN36: First age of use.

df <- read_excel("nyts2022.xlsx") %>%
  dplyr::select(c("newid", "QN1", "QN2", "QN4B", "QN4C", "QN4D",
  "QN4E", "QN5A", "QN5B", "QN5C", "QN5D", "QN5E", "QN133", "QN161",
  "QN162", "QN165", "QN35", "QN36"))
```

```r
# How many individuals?
n_distinct(df$newid) # 28291 individuals

## Consolidate some variables ----

# Hispanic dummy
df <- df %>%
  mutate(hispanic = rowSums(!is.na(dplyr::select(., QN4B:QN4E))), .
    after = QN4E)
df <- df %>%
  mutate(hispanic = case_when(hispanic != 0 ~ "Hispanic", T ~ NA))
df$hispanic <- as.factor(df$hispanic)
df <- subset(df, select = -c(QN4B:QN4E))

## Rename variables for convenience ----
names(df) <- c("id", "age", "male", "hispanic", "native_indian",
"asian", "black", "hawaii_pacific", "white", "social_media_use",
"no_of_cars", "own_bedroom", "school_grades", "cigarette_ever",
"exit_age")

## Coding actual values of variables from codebook ----

# ID
df$id <- as.factor(df$id)

# Age
df <- subset(df, !is.na(age)) # Removes 100 individuals
df <- df %>%
  mutate(age = case_when(age == 1 ~ 9,
                         age == 2 ~ 10,
                         age == 3 ~ 11,
                         age == 4 ~ 12,
                         age == 5 ~ 13,
                         age == 6 ~ 14,
                         age == 7 ~ 15,
                         age == 8 ~ 16,
                         age == 9 ~ 17,
                         age == 10 ~ 18,
                         age == 11 ~ 19, T ~ age))

# Age squared (re-scaled)
df <- df %>%
  mutate(agesq = (age ^ 2) / 100, .after = age)

# Sex
df <- subset(df, !is.na(male)) # Removes additional 169 individuals
df <- df %>%
  mutate(male = case_when(male == 1 ~ 1, T ~ 0))
df$male <- as.factor(df$male)

# Cigarette ever
df <- subset(df, !is.na(cigarette_ever)) # Removes additional 245
    individuals
df <- df %>%
  mutate(cigarette_ever = case_when(cigarette_ever == 1 ~ 1, T ~ 0))

# Age of first cigarette use (or age at interview if never smoked)
df <- df %>%
  mutate(exit_age = case_when(exit_age == 1 ~ 8,
                              exit_age == 2 ~ 9,
                              exit_age == 3 ~ 10,
                              exit_age == 4 ~ 11,
```

```
                                    exit_age == 5 ~ 12,
                                    exit_age == 6 ~ 13,
                                    exit_age == 7 ~ 14,
                                    exit_age == 8 ~ 15,
                                    exit_age == 9 ~ 16,
                                    exit_age == 10 ~ 17,
                                    exit_age == 11 ~ 18,
                                    exit_age == 12 ~ 19, T ~ exit_age)
    )
df <- df %>%
  mutate(exit_age = case_when(is.na(exit_age) ~ age, T ~ exit_age))

# Social media use
df <- subset(df, !is.na(social_media_use)) # Removes additional 1689
    individuals
df$social_media_use <- as.factor(df$social_media_use)
df <- df %>%
  mutate(social_media_use = case_when(social_media_use %in% c(1) ~
  "Never", social_media_use %in% c(2, 3, 4) ~ "Weekly",
  social_media_use %in% c(5, 6, 7, 8) ~ "Daily", T ~ social_media_use
    ))
df$social_media_use <- factor(df$social_media_use, levels = c("Never"
    , "Weekly", "Daily"))

# No. of cars
df <- subset(df, !is.na(no_of_cars)) # Removes additional 1111
    individuals
df$no_of_cars <- as.factor(df$no_of_cars)
df <- df %>%
  mutate(no_of_cars = case_when(no_of_cars == 1 ~ "0",
                                no_of_cars == 2 ~ "1",
                                no_of_cars == 3 ~ ">=2", T ~ no_of_
    cars))
df$no_of_cars <- factor(df$no_of_cars, levels = c("0", "1", ">=2"))

# Own bedroom
df <- subset(df, !is.na(own_bedroom)) # Removes additional 82
    individuals
df <- df %>%
  mutate(own_bedroom = case_when(own_bedroom == 1 ~ 0, T ~ 1))
df$own_bedroom <- as.factor(df$own_bedroom)

# Current school grades
df <- subset(df, !is.na(school_grades)) # Removes additional 197
    individuals
df$school_grades <- as.factor(df$school_grades)
df <- subset(df, !(school_grades %in% c(6, 7))) # Removes additional
    1843 individuals with no information about grades
df <- df %>% mutate(school_grades = case_when(school_grades %in% c(1,
     2) ~ "High", T ~ "Low"))
df$school_grades <- factor(df$school_grades, levels = c("High", "Low"
    ))

## Ethnicity variables ----

# Some more cleaning
df$native_indian <- as.factor(df$native_indian)
df <- df %>%
  mutate(native_indian = case_when(!is.na(native_indian) ~ "Native
    Indian", T ~ native_indian))
df$asian <- as.factor(df$asian)
df <- df %>%
  mutate(asian = case_when(!is.na(asian) ~ "Asian", T ~ asian))
```

```
df$black <- as.factor(df$black)
df <- df %>%
  mutate(black = case_when(!is.na(black) ~ "Black", T ~ black))
df$hawaii_pacific <- as.factor(df$hawaii_pacific)
df <- df %>%
  mutate(hawaii_pacific = case_when(!is.na(hawaii_pacific) ~ "Hawaii/
    Pacific", T ~ hawaii_pacific))
df$white <- as.factor(df$white)
df <- df %>%
  mutate(white = case_when(!is.na(white) ~ "White", T ~ white))

# How many individuals report multiple ethnicity?
df <- df %>%
  mutate(ethnicity = rowSums(!is.na(dplyr::select(., hispanic:white))
    ), .after = white)
count(df, ethnicity)

# Remove the 165 individuals without any information about ethnicity
df <- subset(df, ethnicity != 0)

# Consolidate into one variable
df <- df %>%
  pivot_longer(hispanic:white) %>%
  mutate(ethnicity = ifelse(sum(is.na(value)) == 5, value[!is.na(
    value)], "Mixed"), .by = id, .after = male) %>%
  pivot_wider() %>%
  subset(select = -c(hispanic:white))
df$ethnicity <- factor(df$ethnicity, levels = c("Mixed", "Native
    Indian", "Asian", "Black", "Hawaii/Pacific", "White", "Hispanic"))
```

Next, I plot the survival curves shown in Figure 1 and Figure 2.

```
# Survival Curves ------------------------------------------------

# Survival curve
fit <- survfit(Surv(exit_age, cigarette_ever) ~ 1, df, conf.type = "
    log-log")
autoplot(fit, censor.shape = '+', censor.colour = "black", surv.
    colour = "pink3") +
  theme_classic(base_size = 20) +
  xlab("Age in Years") +
  ylab("Survival function") +
  scale_x_continuous(breaks = seq(0, 50, 1), expand = c(0, 0), limits
    = c(0, 20))

# Survival by sex
fit <- survfit(Surv(exit_age, cigarette_ever) ~ male, df, conf.type =
    "log-log")
autoplot(fit) +
  scale_color_hue(labels = c("Female", "Male")) +
  theme_classic(base_size = 20) +
  guides(fill = "none") +
  theme(legend.position = c(0.2, 0.3)) +
  labs(color = "", x = "Age in Years", y = "Survival function") +
  scale_x_continuous(breaks = seq(0, 50, 1), expand = c(0, 0), limits
    = c(0, 20))

# Smoothed hazard curve (change smoothing parameter lambda as per
    convenience)
fit <- bshazard(Surv(exit_age, cigarette_ever) ~ 1, verbose = F,
    lambda = 100, df)
df_surv <- data.frame(time = fit$time, hazard = fit$hazard,
                      lower.ci = fit$lower.ci, upper.ci = fit$upper.
    ci)
```

```
ggplot(df_surv, aes(time, hazard)) +
  geom_line(color = "purple") +
  geom_ribbon(aes(ymin = lower.ci, ymax = upper.ci), alpha = 0.2,
              fill = "pink3") +
  theme_classic(base_size = 20) +
  xlab("Age in Years") +
  ylab("Hazard") +
  scale_x_continuous(breaks = seq(0, 20, 1), expand = c(0, 0), limits
    = c(0, 20)) +
  scale_y_continuous(breaks = seq(0, 100, 0.01), expand = c(0, 0),
    limits = c(0, 0.06))

# Hazard by sex
as.data.frame.bshazard <- function(x, ...) {
  with(x, data.frame(time, hazard, lower.ci, upper.ci))
}
df_surv <- group_by(df, male) %>%
  do(as.data.frame(bshazard(Surv(exit_age, cigarette_ever) ~ 1, data
    = ., verbose = F, lambda = 100))) %>%
  ungroup()
ggplot(df_surv, aes(x = time, y = hazard, group = male)) + geom_line(
    aes(col = male)) +
  geom_ribbon(aes(ymin = lower.ci, ymax = upper.ci, fill = male),
    alpha = 0.3) +
  labs(color = "", x = "Age in Years", y = "Hazard") +
  theme_classic(base_size = 20) +
  scale_color_hue(labels = c("Female", "Male")) +
  guides(fill = "none") +
  theme(legend.position = c(0.2, 0.9)) +
  scale_x_continuous(breaks = seq(0, 20, 1), expand = c(0, 0), limits
    = c(0, 20)) +
  scale_y_continuous(breaks = seq(0, 100, 0.01), expand = c(0, 0),
    limits = c(0, 0.06))
```

Then I run the Cox models, producing Table 2 and Table 3.

```
# Proportional Hazards Model ---------------------------------------

# Model with all covariates
cox_model_1 <- coxph(Surv(exit_age, cigarette_ever) ~ male +
    ethnicity +
                       social_media_use + no_of_cars + own_bedroom +
                       school_grades, data = df)

# Model summary
summary(cox_model_1)
p.fdr(data.frame(summary(cox_model_1)[["coefficients"]])$Pr...z..) #
    FDR
stargazer(cox_model_1) # Latex table

# Test for proportional hazards
eha::logrank(Surv(exit_age, cigarette_ever), group = male, df) # Male
    , not violated
eha::logrank(Surv(exit_age, cigarette_ever), group = ethnicity, df) #
     Ethnicity, violated
eha::logrank(Surv(exit_age, cigarette_ever), group = social_media_use
    , df) # Social media use, not violated
eha::logrank(Surv(exit_age, cigarette_ever), group = no_of_cars, df)
    # Number of cars, violated
eha::logrank(Surv(exit_age, cigarette_ever), group = own_bedroom, df)
     # Own bedroom, violated
eha::logrank(Surv(exit_age, cigarette_ever), group = school_grades,
    df) # School grades, violated
```

```r
# Model stratifying on variables violating proportional hazards
cox_model_2 <- coxph(Surv(exit_age, cigarette_ever) ~ male + strata(
    ethnicity) +
                     social_media_use + strata(no_of_cars) +
                     strata(own_bedroom) + strata(school_grades),
    data = df)

# Model summary
summary(cox_model_2)
p.fdr(data.frame(summary(cox_model_2)[["coefficients"]])$Pr...z..) #
    FDR
stargazer(cox_model_2) # Latex table
```

Then some additional data cleaning required to implement the split population model.

```r
# Reformat Data for Split Population Model --------------------------

# For row splits, we need id, age, cigarette_ever, and the exit_age
df1 <- df[c("id", "cigarette_ever", "age", "exit_age")]

# Creating the row splits
df1 <- df1 %>%
  mutate(start = 0, end = age) %>%
  dplyr::select(-cigarette_ever) %>%
  gather(cigarette_ever, enter, -id) %>%
  group_by(id) %>%
  arrange(id, enter) %>%
  filter(!is.na(enter)) %>%
  mutate(exit = lead(enter)) %>%
  filter(!is.na(exit), !grepl("time_to_event_out_start", cigarette_
    ever)) %>%
  mutate(event = lead(grepl("time_to_event", cigarette_ever), default
     = 0)) %>%
  dplyr::select(id, enter, exit, event) %>%
  ungroup()

# Cleaning up
df1 <- subset(df1, enter != exit)
df1 <- left_join(df1, dplyr::select(df, id:school_grades), by = "id")
     # Add all columns

# Indicator for cigarette ever
df1$event <- ifelse(df1$exit == df1$age, 0, 1)

# How many people smoke the first time at age of interview?
nrow(subset(df, age == exit_age & cigarette_ever == 1)) # 288
    individuals
id_vector <- subset(df, age == exit_age & cigarette_ever == 1) %>%
  distinct(id) %>%
  pull() # Get the id's of these 288 individuals
df1 <- df1 %>%
  mutate(event = case_when(id %in% id_vector ~ 1, T ~ event))

# Reformat with one year intervals
df1 <- survSplit(df1, cut = c(1:80), start = "enter", end = "exit",
                 event = "event")

# Relocate some columns
df1 <- df1 %>% relocate(c(enter, exit, event), .after = age)
```

Now I run the split population model using the `spduration` package. The following code produces Table 4, and the plots in Figure 3 and Figure 4.

```r
# Split Population Model using `spduration` -------------------------

# Sampling a tenth of the population, to ease computation
id_vector <- df %>%
  distinct(id) %>%
  pull() # Recycling this vector of id's
set.seed(5)
a <- sample(id_vector, n_distinct(id_vector) / 10)
df_model <- subset(df1, id %in% a)

# Variables to capture survival characteristics, needed by `
    spduration` (takes about 45 seconds)
system.time(df_model <- add_duration(df_model, "event", unitID = "id"
    , tID = "exit", freq = "year"))

# Splitting a third of the sample, following Schmidt and Witte (1989)
set.seed(5)
b <- sample(a, n_distinct(a) / 3)
df_train <- subset(df_model, id %in% b) # Training sample
df_test <- subset(df_model, !(id %in% b)) # Test sample

## Log-log model specification 1: All covariates ----
loglog_model_1 <- spdur(
  duration ~ male + ethnicity + social_media_use + no_of_cars + own_
    bedroom + school_grades,
  atrisk ~ male + ethnicity + social_media_use + no_of_cars + own_
    bedroom + school_grades,
  data = df_train, distr = "loglog", silent = T)

# Model summary
summary(loglog_model_1)
p.fdr(loglog_model_1$pval, threshold = 0.05) # FDR
AIC(loglog_model_1) # AIC = 616.31

# Hazard plot
plot(loglog_model_1, type = "hazard", ci = F)

# Latex table
print(xtable(loglog_model_1), type = "latex", comment = F, include.
    rownames = F)

# Prediction on test sample
loglog_test_p <- predict(loglog_model_1, newdata = df_test, na.action
     = na.omit)

# Separation plot
obs_y <- df_test[complete.cases(df_test), "event"]
separationplot(loglog_test_p, obs_y, newplot = F)

## Log-log model specification 2: Conservative ----
loglog_model_2 <- spdur(
  duration ~ male + ethnicity,
  atrisk ~ male + ethnicity,
  data = df_train, distr = "loglog", silent = T)

# Model summary
summary(loglog_model_2)
p.fdr(data.frame(loglog_model_2)[-1, ]$Pr...t.., threshold = 0.05) #
    FDR
AIC(loglog_model_2) # AIC = 649.81

# Hazard plot
plot(loglog_model_2, type = "hazard", ci = F)
```

```r
# Latex table
print(xtable(loglog_model_2), type = "latex", comment = F, include.
    rownames = F)

# Prediction on test sample
loglog_test_p <- predict(loglog_model_2, newdata = df_test, na.action
     = na.omit)

# Separation plot
obs_y <- df_test[complete.cases(df_test), "event"]
separationplot(loglog_test_p, obs_y, newplot = F)
```

Finally, the summary statistics that were in Table 1.

```r
# Summary Statistics ------------------------------------------------

# Full population
as_kable(tbl_summary(df[, 2:11],
                     statistic = list(all_continuous() ~ "{mean} ({
    min}, {max})",
                                      all_categorical() ~ "{n} ({p}%)
    ")), format = "latex")
df %>% filter(cigarette_ever == 1) %>% select(exit_age) %>% summary()
     # Conditional exit age

# Test sample
as_kable(tbl_summary(subset(df, id %in% a)[, 2:11],
          statistic = list(all_continuous() ~ "{mean} ({min}, {max})",
                           all_categorical() ~ "{n} ({p}%)")), format
    = "latex")
subset(df, id %in% a) %>% filter(cigarette_ever == 1) %>% select(exit
    _age) %>% summary() # Conditional exit age
```