Tikrit Journal of Administrative and Economic Sciences, Vol. 20, No. 65, Part (1): 331-347
**Doi: www.doi.org/10.25130/tjaes.20.65.1.19**

**Tikrit Journal of Administrative and Economics Sciences**
**مجلة تكريت للعلوم الإدارية والاقتصادية**
ISSN: 1813-1719 (Print)
E-ISSN: 3006-9149

# Utilizing Cox Regression Analysis and Bootstrapping test for Indicating the Risk of Smoking on Health in Kurdistan

**Nazeera Sedeeq Barznji*, Bekhal Samad Sedeeq,**
**Kawthar Saeed Taha, Duaa Faiz Abullah**
Administration and Economics college/Salahaddin University–Erbil

***Keywords:***
Survival analysis, Cox regression, Bootstrapping, Smoking

**A R T I C L E  I N F O**

***Article history:***
Received 19 Sep. 2023
Accepted 30 Jan. 2024
Available online 31 Mar. 2024



*Corresponding author:
**Nazeera Sedeeq Barznji**
Administration and Economics college/Salahaddin University–Erbil

**Abstract**: Survival analysis is a branch of statistical analysis which analyzing the envisioned of time till one event occurs. it is a widely used technique in biomedical and health services researches. In this study, the data sample takes with size (300) persons from all classes of society in Kurdistan, from the age of (15 years) to (82 years) to analyze Cox regression and bootstrapping test for indicating the risk of smoking for this purpose. The dependent variable which it is the effect of smoking on several diseases represents in the independent (predictor) variables like mouth odor, Lung diseases, Dental disease, Gastrointestinal diseases, Liver diseases, Heart and blood system diseases, Alzheimer, Mental diseases and Cancer. The result of the analyses confirmed that Smoking caused several diseases such as cancer, lung diseases, dental diseases, heart and blood system diseases. For these analyses the IBM SPSS Statistics is a powerful statistical software platform used.

Tikrit Journal of Administrative and Economic Sciences, Vol. 20, No. 65, Part (1): 331-347
**Doi:** www.doi.org/10.25130/tjaes.20.65.1.19

# استخدام تحليل انحدار كوكس واختبار الاقلاع للإشارة الى مخاطر التدخين على الصحة في كوردستان

**نظيرة صديق البرزنجي  بيخال صمد صديق  كوثر سعيد طه  دعاء فائز عبد الله**
**كلية الإدارة والاقتصاد/جامعة صلاح الدين–أربيل**

**المستخلص**

تحليل البقاء هو فرع من التحليل الإحصائي الذي يقوم بتحليل الوقت المتصور حتى وقوع حدث واحد. وهي تقنية تستخدم على نطاق واسع في أبحاث الخدمات الطبية الحيوية والصحية. في هذه الدراسة تم أخذ عينة بيانات بحجم (300) شخص من كافة فئات المجتمع في كردستان من عمر (15 سنة) إلى (82 سنة) لتحليل اختبار انحدار كوكس واختبار الإقلاع للإشارة إلى خطر التدخين على هذه الحالة. الغرض من المتغير التابع وهو تأثير التدخين على عدة أمراض يمثل المتغيرات المستقلة (المنبئية) مثل رائحة الفم، أمراض الرئة، أمراض الأسنان، أمراض الجهاز الهضمي، أمراض الكبد، أمراض القلب والجهاز الدموي، الزهايمر، الأمراض النفسية والسرطان. وقد استخدمت نتيجة التحاليل التي أكدت أن التدخين يسبب عدة أمراض مثل السرطان، أمراض الرئة، أمراض الأسنان، أمراض القلب والجهاز الدموي. لإجراء هذه التحليلات، تعد IBM SPSS Statistics منصة برمجية إحصائية قوية مستخدمة.

**الكلمات المفتاحية:** تحليل البقاء على قيد الحياة، انحدار كوكس، الإقلاع عن التدخين، التدخين.

**Goal:** The goal of Cox proportional hazards regression is to generate a model for the hazard rate of the observed population, which is directly related to the survival function of this population. Since smoking has a major risk for morbidity and mortality so this study related Cox regression for risk of smoking to indicate the type of disease affect smoker

## 1. Introduction

Tobacco smoking is known to be the single largest cause of premature death worldwide. It is the main cause of some types of cancer and cardiovascular diseases. From 2000, research referred to most of smoker in age 11 to 13-years- are smoking two or three cigarettes per a day, Tobacco kills up to half of its users who don't quit *(1-3)*, Tobacco kills more than 8 million people each year, including an estimated 1.3 million non-smokers who are exposed to second-hand smoke. Around 80% of the world's 1.3 billion tobacco users live in low- and middle-income countries.

In 2020, 22.3% of the world's population used tobacco: 36.7% of men and 7.8% of women. In addition, every year, they will spend $1200 or more on tobacco products, to maintain their addiction. Worldwide, smoking will kill five million smokers each year. 20% of adult smokers in United States, addicted with smoking. and since in Europe and Asia the smoking rates %33

of adults it causes to kill 4 out of 10 of smokers. (Ellen T. C. et al., 2020, 189-200) (www.who.int, 2023)

In this paper, we will touch on a type of Survival analysis it involves the modelling of time to event data. Survival analysis is the statistical analysis for the rang of time. ***Cox model*** or Proportional hazards models are a branch of survival models in statistics. between the started time to first event happens, it is defining as a reliability law. Survival models connect the time before some event occurs, to one or more covariates.The baseline hazard function, of the cumulative hazard function represents a generalization of the Nelson-Aalen estimator in non-parametric and if the baseline hazard follows a particular form The Cox model may be specialized in survival analysis. The Cox model is, the semiparametric model and it is comparing hazards between the values of the predictors. (David G. K. & Mitchel Klein, 2012, 10), (Frank Emmert-Streib & Matthias Dehmer, 2019, 4)

**2. Literature review:** Deborah Burr in (1994) this study contains several types of bootstrap confidence intervals for parameters in cox's model, the survival model at fixed time points, and the median survival time best asymptotic method for the regression parameter

Belas, Fiserova E.and Krupickova S. in (2013) apply the Cox regression models when right censoring and delayed entry survival data are considered present study of mitral valve replacement in children under 18 years.

Gongjun X., Bodhisattva S., and Zhiliang Y. in (2014). it shows the consistency of several bootstrap methods for making analysis on a change-point in time in the Cox model with right censored survival data. for the maximum partial likelihood estimation of the change-point.to make a new model.

Minh N. Luu, Minji Han, Tra T. Bui Phuong Thao T. Tran in (2022) Smoking trajectory and cancer risk: A population-based cohort study Smoking, even at low doses, increases the risk of most cancers in men. Quitting or reducing smoking, especially at a young age, can lower cancer incidence and mortality. This study may provide more objective results on the relationship between smoking and cancer, because smoking behavior was examined at multiple time points.

**Doi: www.doi.org/10.25130/tjaes.20.65.1.19**

## 3. Theoretical Aspect

**3-1. Survival Analysis:** Survival analysis is interested in studying the time between entry to explore and a review the event. Survival operation explain a life extension from a special starting time to the first event. In the literature of survival analysis, time at the appearance of an event is regarded as a random variable, referred to as event time, failure time, or survival time.

The cox model or the proportional hazard model is a branch of survival models in statistics. Survival models connect the past time before event occurs and the covariates that may be related with that deal of time.

**Survival distribution function and Life Function:** Survival data are generally explaining of the two probability models, survival and hazard. The survival probability (survivor function) S(t) is the probability that an individual survives from the starting time to a specified future time t. It is essential to a survival analysis because survival probabilities for different values of supply conclusive summary datum from time to event data.

The distribution function defined as T can be described as a non-negative random variable.

$$F(t) = P(T \leq t): t \geq 0 \qquad \dots (1)$$

This function gives the probability that the life time of an element in the community will not exceed time t This function has the probability that the element will die in time [0, t)

can also describe the life time T by life function defined by

$$S(t) = P\{T > t\} = 1 - F(t) \quad .t > 0 \qquad \dots (2)$$

This function gives the probability that the life time of an element in the community will exceed time t This function has the probability that the element will not die in time [0, t)

We also use the density function to describe the life time, which is the derivative of the distribution function:

$$f(t) = F'(t) = -S'(t) \qquad \dots (3)$$

Since T a non-negative continuous random variable, we can write as

$$F(t) = pr(T > t) = \int_0^1 f(u)du \qquad \dots (4)$$

$$S(t) = pr(T > t) = \int_1^x f(u)du = 1 - F(T) \qquad \dots (5)$$

(Frank Emmert-Streib & Matthias Dehmer, 2019, 4)

The most common in life

It is noted that the life function has the following properties

**Doi:** www.doi.org/10.25130/tjaes.20.65.1.19

*1) It is incremental function for every t > 0*

$S(t) > S(t+\propto) \quad \forall \propto > 0$ … (6)

Which $S(t) = 1 - F(t)$ from this result by taking the first derivative

$\frac{dS(t)}{dt} = -f(t) \leq 0 : t > 0$ … (7)

At zero moment, all the studied elements are alive, and this is expressed by the relationship:

$S(0) = 1$

When time ends to infinity, all elements will die, (the studied event is realized)

$\lim_{n\to\infty} S(t) = 0$ … (8)

(David G. K. & Mitchel Klein, 2012, 10), (Frank Emmert-Streib & Matthias Dehmer, 2019, 4)

**3-1-1. Cumulative Survival**

$X_0$ Is the Survival for an individual

The cumulative survival=$S(T:X) = exp\left(-C(T:X)\right)$ … (9)

$$exp\left(-C_0(T)^{e^{\sum_{i=1}^{P} X_i b_i}}\right) = \left[e^{-C_0(T)}\right]^{e^{\sum_{i=1}^{P} X_i b_i}} \quad \ldots (10)$$

(Al-kredi, & Altengi, 2014, 2), (Gongjun X., et al., 2014, 6)

**We note from the function the following properties**:

1. Always non-negative (wavelength).
2. It has no upper bound for the two functions (the survival function, the risk function), the survival state is the most natural when Data analysis because it directly describes the survival evolution of a set of data.

As the advantages of the risk function are:

1. Measures the immediate possibility, as the survival function is a cumulative measure over time.
2. It is used to identify the shape of the form.
3. The risk function is the method that helps us to find the mathematical model for survival data

The survival model is written as limits for the risk function.

**3-2. Cox Regression Model:** David Cox is an English statistician, and are known one at that. He has written over 300 papers or books on a variety of topics, has advised government, was knighted for his contribution to science,

and holds numerous fellowships and awards. His paper introducing the proportional hazards assumption and inference for it (Cox, 1972),

Cox Regression is a famous regression technique for survival outcomes and it is an expected model for time-to-event data. (John F., 2008: 4)

**3-2-1. Cox Proportional hazards method [CPHM]:** It has a very important effect in survival analysis. in survival's regression model, The Cox proportional hazards model, is the most special applied, it explores the linked between predicted variable and the time-to-event over the hazard function. It refers the effect of predictors on the hazard and this effect is constant over time, i.e.,

$C(t|X) = C_0(t) \exp(b_1X_1 + b_2X_2 + \cdots + b_PX_P)$

where C (t |X) is the hazard at time t for predictors $X_1,\ldots, X_P$, $C_0(t)$ is the basic hazard function, and $b_1,\ldots, b_p$ are the parameters they are the effect of predictors on the overall hazard.

The hazard is generally denoted by C(t) is the probability of person who is under observation at a time t has an event at that time, it shows that the quick event rate for a person who has already survived to time t.

$$C(t|X) = C_0(t)\exp(b_1X_1 + b_2X_2 + \cdots + b_PX_P) \quad \ldots (11)$$

Where C (t) is the predicted hazard at time t, (Al-kredi, & Altengi, 2014: 2), (Bela S., et al., 2013: 22), (Brandon G., et al., 2014, 4), (Efron, B., 1977, 4)

$C_0(t)$ represents the hazard when all of the predictors $X_1, X_2 \ldots, X_p$ are equal to zero.

Consider a simple model with one predictor, $X_1$, a special case of the Cox proportional hazards: because not included $b_0$.

by generalize to P covariates $C(t|X) = C_0(t)exp\left(\sum_i^p b_iX_i\right)$ … (12)

$$\frac{C(t/X)}{C_0(t)} = exp\left(\sum_i^p b_iX_i\right) \quad \ldots (13)$$

$$C_m(t|X) = C_0(t)\, exp^{\sum_i b_iX_{mi}} \quad \ldots (14)$$

$$C_n(t|X) = C_0(t)\, exp^{\sum_i b_iX_{ni}} \quad \ldots (15)$$

$$\frac{C_m(t|X)}{C_n(t)} = \frac{C_0(t)\, exp^{\sum_i b_iX_{mi}}}{C_0(t)\, exp^{\sum_i b_iX_{ni}}} = exp^{\sum_I b_i(X_{mi}-X_{ni})} \quad \ldots (16)$$

**Doi:** www.doi.org/10.25130/tjaes.20.65.1.19

$$CR = exp^{\sum_I b_i(X_{mi}-X_{ni})} \quad \dots (17)$$

$$C_m(t|\mathrm{X}) = CR\ X\ C_n(t|\mathrm{X}) \quad \dots (18)$$

(Al-kredi, & Altengi, 2014, 2), (Mark, S., 2007, 23)

**3-2-2. Cumulative Hazard:**

$$C(t|\mathrm{X}) = \int_0^T C(u.X)du \quad \dots (19)$$

$$= \int_0^T C_0(u)e^{\sum_{i=1}^P x_i b_i}\, du$$

$$= e^{\sum_{i=1}^P X_i b_i} \int_0^T C_0(u)\, du$$

$$= C_0(T)\ e^{\sum_{i=1}^P X_i b_i} \quad \dots (20)$$

$$\hat{S}(t|X_O) = \hat{S}_O(t)\exp^{\sum_{i=1}^P X_{io}\,\hat{b}_i} \quad \dots (22)$$

By Survival Cumulative, we estimate odd Cumulative with Covariate of $X_o$

$$\hat{S}(t|X_O) = \hat{S}_O(t)^{\exp\sum_{i=1}^P X_{io}\,\hat{\beta}_i} \quad \dots (23)$$

(John F., 2008: 5), (Al-kredi, & Altengi, 2014: 2), (Mark, S., 2007: 23)

**3-2-3. Parameters estimation of the Cox regression model:** One of the conditions for finding the maximum possible function is that the distribution must be known. The probability function is defined as the product of the common function of all observations within the sample under the hypothesis given. If the Cox model is used then the t-distribution is known, the maximum possibility function (b) cannot be found. In the laboratory models, while in the semi-parametric Cox model, we find the possible function based about event (death) rather than the death event distribution. (Mark, S., 2007, 22), (Brandon G., et al., 2014, 3)

**3-2-4. The Log Likelihood Significance test of the model as a whole (Maximum likelihood-ratio Test)**

Let $t = 1.2. \dots . M$ are M unique failure times $T_1 . T_2 . \dots . T_N$ that occur at time $T_t$ is referred to as $D_i$. Let c and d= l, …, m,. all persons that are at risk instantly before time $T_t$ is represent as $R_t$. is called the risk set includes all individuals that fail at time $T_t$ as well as those that are censored or fail at a time later than $T_t$ Let r=1,…, n index the members of $R_t$.Let $X$ to asset of p covariates. The covariates are indexed by subscripts $i.j\ or\ k$.The values of

**Doi:** www.doi.org/10.25130/tjaes.20.65.1.19

the covariates at particular failure time $T_d$ are written $X_{1d}.X_{2d}.X_{3d}.\ldots.X_{td}$ in general the regression coefficient to estimated are $b_1\,.\,b_2\,.\,b_3\,.\ldots.\,b_K$

The model test is an important test, which model is suitable using the test of the maximum likelihood-ratio Where this test is carried out by estimating the two models for the data and then comparing the first model with the second model from value better. Must find out that this difference is significant using the formula:

$$LR = -2log(L_m L_\circ) = -2[logL_\circ + logL_m] \quad \ldots (24)$$

$L_\circ$: is the maximum likelihood function with only the basic risk function and that the test statistic result has the chi-square distribution

If there no ties among the failure times the log likelihood is given as follow:

$$L(b) = \sum_{t=1}^{N}\left\{\left(\sum_{I=1}^{K} X_t\, b_i\right) - ln\sum_{r\in R_i} exp\left(\sum_{I=1}^{K} x_{\ tr}\, b_i\right)\right\}$$

$$= \sum_{t=1}^{N}\left\{\left(\sum_{I=1}^{K} X_{it}\, b_i\right) - ln\left(G_{R_i}\right)\right\} \quad \ldots (25)$$

$$G_R = \sum_{r\in R} exp\left(\sum_{I=1}^{K} X_{\ tr}\, \beta_i\right) \quad \ldots (26)$$

By taking first and second derivations

$$H_{jR} = \frac{\partial G_R}{\partial \beta_{ij}} = \sum_{r\in R} X_K\left(\sum_{I=1}^{K} X_{\ tr}\, b_i\right) \quad \ldots (27)$$

$$A_{jR} = \frac{\partial^2 G_R}{\partial \beta_j \partial \beta_M} = \frac{\partial H_{jR}}{\partial \beta_M} = \sum_{r\in R} X_K\left(\sum_{I=1}^{K} X_{\ tr}\, b_i\right) \quad \ldots (28)$$

$$= \sum_{r\in R} X_{jr} x_{mr}\left(\sum_{I=1}^{K} X_{\ tr}\, b_i\right)$$

Then maximum likelihood solution is found by Newton a Raphson this method requires the first and second order partial derivatives. The first order partial derivatives are

$$U_j = \frac{\partial LL(b)}{\partial \beta_j} = \frac{\partial}{\partial \beta_M} = \sum_{I=1}^{K}\left\{X_{\ jt} - \frac{H_{jR}}{G_R}\right\} \quad \ldots (29)$$

The second order partial derivatives

$$I_{jM} = \sum_{I=1}^{K}\left\{A_{\ jMR} - \frac{H_{jR}H_{MR}}{G_R}\right\} \quad \ldots (30)$$ (Mark, S., 2007, 7), (Gongjun X., et al., 2014, 6)

### 3.2. 5. Standard Errors of Survival Probabilities:

$$S_E(S_t) = S_t\sqrt{\sum \frac{D_t}{R_t(R_t - D_t)}} \quad \ldots (31)$$

Where: $\frac{D_t}{R_t(R_t - D_t)}$ is a risk $(R_t)$

**Doi:** www.doi.org/10.25130/tjaes.20.65.1.19

($D_t$): Standard errors,$S_E(S_t)$: is the margin of error used the 95 % confidence interval estimates. (Toshinari, K., 2004, 8), (Rebecca Z. & Jeffrey C. S., 2005, 8)

**3.2.6. Tests of hypotheses:**

**Wald statistics.**

$$\textbf{Wald statistics} = \left(\frac{\boldsymbol{Estimate\ of\ b_i}}{\boldsymbol{SE(b)}}\right)^2 \quad \dots (32)$$

This means that testing the effect of the predictor's variables on the stay time in the model

**3.2.7. Cox Snell Residuals**

The Cox Snell Residuals defined as

$$H_{bO}\ (K_t) = \sum_{T_i \leq T_t} \left[\frac{N_i}{\sum_{j=R_{K_t}} S_t}\right] \quad \dots (33)$$

Where: $B's$: regression coefficients, $H_{bO}\ (K_t)$: cumulative baseline hazard function. (Mitchell H. et al., 1993: 708), (Al-Kredi, & Altengi, 2014, 5), (Patrick Breheny, 2010: 5)

**3.2.8. Goodness of fit tests:**

**Chi Square Test**

$$\mathcal{X}^2 = \frac{(\sum O_{1t} - \sum E_{1t})}{\sum var(E_{1t})} \quad \dots (34) \quad \text{where } var(E_{1t}) = \frac{N_{1t}N_{2t}\,(N_t - O_t)}{N_t\,(N_t - 1)}$$

Where $\sum O_{jt}$ the sum of the **observed number of events** in the $j^{th}$ group $\sum E_{jt}$: the sum of the expected number of events in the $j^{th}$ group over time. (Mark, S., 2007, 20)

**3-2-9. Bootstrapping:** Bootstrap methods: another looks at the jackknife Bootstrap (Efron, 1996) is one of the several ways how to do re-sampling. In bootstrap we approximate the entire sampling distribution by re-sampling original data sets. It is useful when the original sample is small and the assumption of normality does not hold. Here we use bootstrap for the estimation of regression parameters and their standard errors and for the determination of the confidence interval It methods used in estimation of the sampling distribution, It determines the measures of accuracy to sample estimates.; It may be also used for making hypothesis tests. (Bela, S., et al., 2013: 27)

**3-3. Baseline Hazard:** Function which depends on time and analogy for the explanatory variables vector = 0 $_i X.b$: $is\ a\ vertical\ vector\ 1 \times P$ of the

**Doi:** www.doi.org/10.25130/tjaes.20.65.1.19

unknown features and uses the partial potential method to estimate the features unknowns.$exp\ (bX_i)'$: is the relative risk that does not depend on time, that is, the effect of the explanatory variables is an increase or the decrease in risk is constant and does not change depending on the change in time point t, and also the ratio between two rates of risk.

$S_0(T)$ is cumulative hazard Function using

$$S_0(T_0) = \prod_{T_1<T_0} \propto_t$$

$$\propto_t = \frac{S(T_t)}{S(T_{t-1})} = \left[\frac{S_0(T_t)}{S_0(T_{t-1})}\right]^{e^{\sum_{i=1}^{P} X_i b_i}} = \left[\frac{S_0(T_t)}{S_0(T_{t-1})}\right]^{\theta_T} \quad \dots (35)$$

Where $\theta_T = e^{\sum_{i=1}^{P} X_i b_i}$

(Hamad, F. et al., 2022: 239)

## 4. Practical Aspect

**4-1. Data Description:** The data sample consists (300) persons from all classes of society, from the age of (15 years) to (82 years),

**4-2. Variable Description:**

$y_i = Dependent\ (response)$

$X_i = Independent(Explanatory)\ Variables =$

$X_1$ = Mouth odor, $X_2$ = Lung diseases, $X_3$ = Dental disease, $X_4$ = Gastrointestinal diseases, $X_5$ = Liver diseases, $X_6$ = Heart and blood system diseases, $X_7$ = Alzheimer, $X_8$ = Mental diseases $X_9$ = Cancer.

4-3. Cox Regression Analysis

Table (1): Summary Cox Regression Analysis

| Cases | Numbers | Ratios |
|---|---|---|
| Event | 285 | 95% |
| Censored | 15 | 5% |
| Total | 300 | 100% |

Table (1) shows 300 persons which (285) are events or persons are smoking with (95 %), and (15) persons are Censored with 5%

Table (2): gender (females and male)

| Stratum | Event | Censored | Censored Percent |
|---|---|---|---|
| Females | 2 | 0 | 0.0% |
| males | 283 | 15 | 5.0% |
| Total | 285 | 15 | 5.0% |

**Doi:** www.doi.org/10.25130/tjaes.20.65.1.19

Table (2) shows 300 persons by gender which females are (2) events or persons are smoking cigarettes, (0) Censored with (0.0%) Censored Percent and Males are (283) events with and (15) persons are Censored with (5.0%)

**4-3-1. Assessing the model Stepwise Forward Selection Initial step (Beginning Block) Initial step (Beginning Block)** is a model without predictor variables, a model with the predictor variables set to 0.

Table (3): displays the -2-log likelihood

| -2 Log Likelihood | Cox & Snell R Square |
|---|---|
| 2707.914 | 0.717 |

Table (3) displays the -2 log likelihood (2707.914) and Cox & Snell R Square (0.717) is predict the changing in the explanatory variables it indicates that the risk of smoking on health by the ratio (71.7)

**4-3-2. Step One:**

Table (4): Test of Model Coefficients by Omnibus Test

| -2Log Likelihood | score | | | Change in Step | | | Change in Block | | |
|---|---|---|---|---|---|---|---|---|---|
| | Chi-square | df | P.V. | Chi-square | df | P.V. | Chi-square | df | Sig. |
| 2675.160 | 26.901 | 9 | .021 | 32.753 | 9 | 0. 020 | 32.753 | 9 | 0.020 |

Comparing Table (4) and Table (3) by -2 log likelihood reduced from 2707.914in Table (3) to 2675.160 Table (4), but an improvement of chi-square statistic with nine degrees of freedom form (26.901) to (32.753) with both significant (p-value = 0.020) it indicates the existence of risk of smoking on health.

**4-3-3. Variables in the Equation**

Table (5): Variables in the Equation with the coefficients (B)

| | B | SE | Wald | df | Sig (p value) | Exp(B) |
|---|---|---|---|---|---|---|
| th odor =$X_1$ | 0.017 | 0.138 | 0.015 | 1 | 0.902 | 1.0171 |
| Lung diseases =$X_2$ | 0.167 | 0.075 | 4.918 | 1 | 0.027 | 1.1817 |
| Dental disease= $X_3$ | 0.122 | 0.057 | 4.550 | 1 | 0.033 | 1.1297 |
| Gastrointestinal diseases =$X_4$ | 0.019 | 0.064 | 0.084 | 1 | 0.772 | 1.0191 |
| Liver diseases=$X_5$ | 0.728 | 0.719 | 1.058 | 1 | 0.311 | 2.072 |
| Heart and blood system diseases= $X_6$ | 0.182 | 0.062 | 2.9354 | 1 | 0.003 | 1.1996 |
| Alzheimer= $X_7$ | 0.014 | 0.035 | 0.138 | 1 | 0.710 | 1.014 |
| Mental diseases =$X_8$ | 0.000 | 0.027 | 0.000 | 1 | 0.993 | 1.000 |
| Cancer =$X_9$ | 0.306 | 0.125 | 5.992 | 1 | 0.015 | 1.3579 |

**Doi: www.doi.org/10.25130/tjaes.20.65.1.19**

Table (5) shows the variables in the Equation by (p-value smaller than 0.05) emphasize to the risk of smoking on health cause to the following diseases

[$X_6$ = Heart and blood diseases with p-value = $_{0.003}$ has, $X_9$ = Cancer with p-value = (0.015), $X_2$ = Lung diseases with p-value = (0.027) and $X_3$ = Dental disease with p-value = (0.033)].

**4-4. Bootstrap used to identify the accuracy:**

Table (6): Bootstrap Specifications

| | |
|---|---|
| Sampling Method | Simple |
| Number of Samples | 1000 |
| Confidence Interval Level | 95.0% |
| Confidence Interval Type | Percentile |

**4-5. Bootstrap for Variables in the Equation:**

Table (7): Bootstrap for Variables in the Equation

| | **B** | **Bootstrap** | | |
|---|---|---|---|---|
| | | **Bias** | **Std. Error** | **Sig (p-value)** |
| Mouth odor =$X_1$ | 0.013 | 0.001 | 0.126 | 0.905 |
| Lung diseases =$X_2$ | 0.169 | 0.000 | .075 | 0.018 |
| Dental disease= $X_3$ | 0.121 | 0.006 | .051 | 0.007 |
| Gastrointestinal diseases =$X_4$ | 0.019 | 0.004 | .057 | 0.735 |
| Liver diseases=$X_5$ | 0.737 | .637 | 2.104 | .336 |
| Heart and blood diseases= $X_6$ | 0.182 | -0.001 | .078 | .024 |
| Alzheimer= $X_7$ | 0.014 | 0.000 | 0.030 | .638 |
| Mental diseases =$X_8$ | 0.001 | 0.002 | 0.023 | .978 |
| Cancer =$X_9$ | 0.306 | -0.365 | 0.492 | .009 |

In Table (7) by the Bootstrap for variables in the Equation by sig (p-value) emphasize that the risk of smoking are only on the same four disease [ $X_3$ = Dental disease (0.007), $X_9$ = Cancer with (0.009), $X_2$ Lung diseases (0.018) and $X_6$ = Heart and blood diseases (0.024) it indicate to the accuracy the of regression statistical analysis.

**4-6. Baseline Hazard Function:**

**Doi: www.doi.org/10.25130/tjaes.20.65.1.19**

Table (8): Illustrates the Baseline of Hazard Function

| Baseline Survival Table | | | | | |
|---|---|---|---|---|---|
| **y** | **Time (age)** | **Baseline Cum Hazard** | **At mean of covariates** | | |
| | | | **Survival** | **SE** | **Cum Hazard** |
| y = 0 | 36 | 0.813 | 0.516 | 0.247 | 0.662 |
| y = 1 | 15 | 0.004 | 0.997 | 0.003 | 0.003 |
| | 16 | 0.020 | 0.984 | 0.007 | 0.016 |
| | 18 | 0.040 | 0.968 | 0.010 | 0.033 |
| | 19 | 0.049 | 0.961 | 0.011 | 0.040 |
| | 20 | 0.091 | 0.928 | .015 | 0.074 |
| | 21 | 0.126 | 0.902 | .017 | 0.103 |
| | 22 | 0.168 | 0.872 | .019 | 0.136 |
| | 23 | 0.196 | .853 | .020 | 0.159 |
| | 24 | 0.225 | .833 | .021 | 0.183 |
| | 25 | .255 | .813 | .022 | 0.207 |
| | 26 | .260 | .809 | .022 | 0.211 |
| | 27 | .285 | .793 | .023 | 0.232 |
| | 28 | .311 | .776 | .024 | 0.254 |
| | 29 | .322 | .769 | .024 | .262 |
| | 30 | .360 | .746 | .025 | .293 |
| | 31 | .382 | .733 | .025 | .311 |
| | 32 | .411 | .716 | .026 | .334 |
| | 33 | .439 | .699 | .026 | .358 |
| | 34 | .469 | .683 | .027 | .382 |
| | 35 | .506 | .663 | .027 | .412 |
| | 36 | .531 | .649 | .027 | .432 |
| | 37 | .550 | .639 | .028 | .448 |
| | 38 | .570 | .629 | .028 | .464 |
| | 39 | .590 | .619 | .028 | .480 |
| | 40 | .617 | .605 | .028 | .502 |
| | 41 | .651 | .588 | .028 | .530 |
| | 42 | .710 | .561 | .029 | .578 |
| | 43 | .748 | .544 | .029 | .609 |
| | 44 | .764 | .537 | .029 | .622 |
| | 45 | .829 | .509 | .029 | .675 |
| | 46 | .871 | .492 | .029 | .710 |

**Doi: www.doi.org/10.25130/tjaes.20.65.1.19**

| Baseline Survival Table | | | | | |
|---|---|---|---|---|---|
| y | Time (age) | Baseline Cum Hazard | At mean of covariates | | |
| | | | Survival | SE | Cum Hazard |
| | 47 | .889 | .485 | .029 | .724 |
| | 48 | .943 | .464 | .029 | .768 |
| | 49 | .972 | .453 | .029 | .791 |
| | 50 | 1.082 | .414 | .029 | .881 |
| | 51 | 1.125 | .400 | .028 | .916 |
| | 52 | 1.194 | .378 | .028 | .972 |
| | 53 | 1.241 | .364 | .028 | 1.011 |
| | 54 | 1.279 | .353 | .028 | 1.042 |
| | 55 | 1.358 | .331 | .027 | 1.106 |
| | 56 | 1.444 | .308 | .027 | 1.176 |
| | 57 | 1.539 | .286 | .026 | 1.253 |
| | 58 | 1.572 | .278 | .026 | 1.280 |
| | 60 | 1.717 | .247 | .025 | 1.398 |
| | 61 | 1.776 | .235 | .025 | 1.446 |
| | 62 | 1.861 | .220 | .024 | 1.515 |
| | 63 | 2.049 | .188 | .023 | 1.669 |
| | 64 | 2.128 | .177 | .022 | 1.733 |
| | 65 | 2.277 | .157 | .021 | 1.854 |
| | 66 | 2.341 | .149 | .021 | 1.907 |
| | 67 | 2.481 | .133 | .020 | 2.021 |
| | 68 | 2.558 | .125 | .020 | 2.083 |
| | 69 | 2.775 | .104 | .018 | 2.260 |
| | 70 | 3.318 | .067 | .014 | 2.702 |
| | 71 | 3.399 | .063 | .014 | 2.767 |
| | 72 | 3.679 | .050 | .012 | 2.996 |
| | 73 | 3.788 | .046 | .012 | 3.085 |
| | 74 | 4.042 | .037 | .011 | 3.291 |
| | 75 | 4.552 | .025 | .009 | 3.707 |
| | 78 | 5.038 | .017 | .007 | 4.102 |
| | 79 | 5.752 | .009 | .005 | 4.684 |
| | 80 | 7.134 | .003 | .002 | 5.809 |
| | 82 | 9.102 | .001 | .001 | 7.412 |

**Doi:** www.doi.org/10.25130/tjaes.20.65.1.19

Table (8) the Baseline Hazard Function which depends on time and analogy for the explanatory variables vector = 0.1 it shows that the increasing in smoking time the hazard is also increases.

## 5. Conclusion and *Recommendation*

**5-1. Conclusion:** In the practical aspect the results:

1. Shows 300 persons which (285) are events or persons are smoking cigarettes with (88.2%), and (15) persons are Censored with 4.6 % shows 300 persons by gender which females are (2) event or persons are smoking (0) Censored with (0.0%) Censored Percent and Males are (283) events with, and (15) persons are Censored with (5.0%)
2. In Omnibus Tests of Model Coefficients by Cox & Snell R Square It predicts the effect of smoking appears with ratio (%71.7).
3. Comparing with -2 log likelihood reduced from (2707.914) in Table (3) to (2675.160) in table (4), but an improvement of chi-square statistic with nine degrees of freedom form (26.901) to (32.753) with both significant (p-value = 0.020) it indicates the existence of risk of smoking on health.
4. In the variables in the Equation by the Sig. (p-value) of the coefficient, and the Wald statistic emphasize on the risk of smoking on health only cause to following diseases [X6 = Heart and blood diseases = 0.003 has, X9 = Cancer (0.015), X2 = lung diseases (0.027), and X3 = Dental disease (0.033).
5. By the Bootstrap for variables in the Equation by sig (p- value) emphasize that the risk of smoking is only on the same four disease [ X3 = Dental disease (0.007), X9 = Cancer with (0.009), X2 = lung diseases (0.018), and X6 = Heart and blood diseases (0.024) it indicate to the accuracy the of regression statistical analysis.

### 5.2. Recommendation

1. It is considered one of the suitable models for binary data, through which the survival time of the patient is, studied the researcher recommends that it used in various fields in medical domains.
2. It is considered one of the modern methods that are characterized by the accuracy of its results, the researcher recommends that it used in various scientific fields.

## References

1. Agathe Guilloux Professeure au LaMME, 2022, "Survival and longitudinal data analysis Chapter 2: tests and the Cox model"- Université d'Évry - Paris Saclay.

2. Bela S., Fiserova E. and Krupickova S., 2013, "Study of Bootstrap Estimates in Cox Regression Model with Delayed Entry", Acta Universitatis Paluckianae Olemucensis, Facultas Rerum Naturalium. Mathematica, Volume 52, Issue 2.
3. Brandon G., Samantha S., and Inmaculada A., 2014, "Survival analysis and regression models", Journal of Nuclear Cardiology, Volume 21, pp. 686-694, Published online 2014 May 9. Doi:10.1007/s12350-014-9908-2.
4. David G. Kleinbaum, and Mitchel Klein, 2012, "Survival Analysis A Self-Learning Text", Third Edition, For further volumes: http://www.springer.com/series/2848.
5. Al-Kredi Khudr and Altengi Maen, 2014, "Finding the Least Possible Hazards in Cox Regression Model", PhD. Thesis, DOI: 10.13140/2.1.1273.4244.
6. Efron, Bradley, 1977, "The Efficiency of Cox's Likelihood Function for Censored Data", Journal of the American Statistical Association, Vol. 72, No. 359, pp. 557-565, DOI:10.1080/01621459.1977.10480613.JSTOR 2286217.
7. Ellen T. Chang, Edmund C. Lau & Suresh H. Moolgavkar, 2020, "Smoking, air pollution, and lung cancer risk in the Nurses' Health Study cohort: time-dependent confounding and effect modification", National Library of Medicine, National Center for Biotechnology, Vol. 50, No. 3, pp. 189-200, Published online: https://doi.org/10.1080/10408444.2020.1727410.
8. Hamad Farag, Abdulkarim Salem, Hamad Ayman, 2022, "Mixture Method to Estimate Baseline Hazard for Non-Arbitrary Function of the Cox Proportional Model", International Journal of Sciences: Basic and Applied Research (IJSBAR), Vol. 62, No. 2, pp. 235-248, ISSN 2307-4531(Print & Online).
9. Frank Emmert-Streib, and Matthias Dehmer, 2019, "Introduction to Survival Analysis in Practice", Machine Learning and Knowledge Extraction, Vol. 1, No. 3, pp. 1013-1038, https://doi.org/10.3390/make1030058.
10. Gongjun X., Bodhisattva S., and Zhiliang Y., 2014, "Bootstrapping a change-point Cox model for survival data". Electronic journal of statistics, School of Statistics, University of Minnesota, Vol. 8, ISSN: 1935-7524.
11. John F., 2008, "Cox Proportional-Hazards Regression for Survival Data", Appendix to An R and S-PLUS Companion to Applied Regression 15 June 2008.
12. Mitchell H., Katz MD., and Walter, W. Hauck, 1993, "Proportional Hazards (Cox) Regression", Journal of General Internal Medicine, Vol. 8, Dec.,1993.
13. Patrick Breheny, 2010, "Residuals and model diagnostics", https://myweb.uiowa.edu/pbreheny/7210/f17/notes/11-07.pdf.
14. Mark Stevenson, 2007, "An Introduction to Survival Analysis", Epi Centre, IVABS, Massey University, December 2007, http://www.biecek.pl/statystykaMedyczna/Stevenson_survival_analysis_195.721.pdf.
15. Rasmus Waage Petersen, 2022, "Cox's proportional hazards / regression model - model assessment", https://people.math.aau.dk/~rw/Undervisning/DurationAnalysis/Slides/lektion4.pdf.
16. Rebecca Zwick and Jeffrey C. Sklar, 2005, "A Note on Standard Errors for Survival Curves in Discrete-Time Survival Analysis", Journal of Educational and Behavioral

Statistics, Vol. 30, No. 1 (Spring, 2005), pp. 75-92., Published By: American Educational Research Association.

17. Toshinari, Kamakura, 2004, "Computational Methods in Survival Analysis" Papers, No. 2004,29, Humboldt-Universität zu Berlin, Center for Applied Statistics and Economics (CASE), Berlin.
18. Wenbin Lu, 2008, "Maximum Likelihood Estimation in the Proportional Hazards Model", Ann Inst. Stat. Math., 60: 545–574, DOI 10.1007/s10463-007-0120.
19. West, R., Shiffman, S., 2007, "Fast Facts: Smoking Cessation", Health Press Ltd., p. 28., ISBN 978-1-903734-98-87.
20. https://www.who.int/news-room/fact-sheets/detail/tobacco July 2023.