Tikrit Journal of Administrative and Economic Sciences, Vol. 21, No. 71, Part (1): 512-527
Doi: www.doi.org/10.25130/tjaes.21.71.1.28

Tikrit Journal of Administrative and Economics Sciences
مجلة تكريت للعلوم الإدارية والاقتصادية
EISSN: 3006-9149
PISSN: 1813-1719

The Impact of Lifestyle and Demographic Factors on thyroid Cancer Incidence Using a Probit Regression Model

Sawen Othman Babakr*, Sami Ali Obed, Zhean Muhsein Jalal
College of Management and Economics/Salahaddin University

Keywords:
Regression, probit analysis, Thyroid cancer, Demographic variables, bivariate probit model, ATPO.

A R T I C L E   I N F O

Article history:

Received	12 Nov. 2024
Received in revised form	21 Dec. 2024
Accepted	20 Jan. 2025
Available online	30 Sep. 2025

*Corresponding author:
Sawen Othman Babakr
College of Management and Economics/Salahaddin University

Abstract: This research studies the impact of demographic and lifestyle variables on the occurrence of thyroid cancer. The data has been collected from Smart Health Tower in Sulaimani, encompassing 100 instances of cancer and thyroid disorders. The Probit regression analysis incorporated the subsequent covariates. These parameters included in the model age, gender, tumor size, multifocality, vascular invasion, capsular invasion (CI), extrathyroidal extension (ETE), involvement of lymph nodes (LN), and distant metastases. In order to create a bivariate probit model, researchers often characterize the joint distribution of the unobservable as a bivariate normal distribution.

The Probit model was employed to investigate the relationship between these parameters and cancer incidence for risk estimation. The results specify that certain factors, such as tumor laterality and thyroid-stimulating hormone (TSH) levels, were deemed less significant, but others, including gender, lymph node invasion, and preoperative ATPO levels, exhibited a robust association with an increased cancer risk. The results indicated that gender, lymph node invasion, and preoperative ATPO levels are substantially correlated with an elevated risk of thyroid cancer.

Although certain parameters, such as TSH levels and tumor laterality, were shown to be less significant, the results demonstrate robust correlations between cancer incidence and other covariates. The findings underscore the imperative for holistic models to evaluate the diverse interacting elements in cancer risk assessment.

Tikrit Journal of Administrative and Economic Sciences, Vol. 21, No. 71, Part (1): 512-527
Doi: www.doi.org/10.25130/tjaes.21.71.1.28

تأثير نمط الحياة والعوامل الديموغرافية على الإصابة بالسرطان الغدة الدرقية باستخدام نموذج الانحدار بروبيت (probit)

ساوين عثمان بابكر　　　　سامي علي عبيد　　　　ژيان محسن جلال
كلية الادارة و الاقتصاد/جامعة صلاح الدين-اربيل

المستخلص

تبحث هذه الدراسة في تأثير نمط الحياة والعوامل الديموغرافية على الإصابة بالسرطان باستخدام نموذج الانحدار بروبيت. من خلال التركيز على متغيرات مثل العمر والجنس وخصائص الورم والعلامات الحيوية قبل الجراحة، يهدف البحث إلى تحديد التنبؤات المهمة لخطر الإصابة بالسرطان، وخاصة سرطان الغدة الدرقية والجهاز الهضمي. يشير التحليل الإحصائي إلى أن عوامل مثل مستويات ATPO قبل الجراحة غزو العقد الليمفاوية تؤثر بشكل كبير على خطر الإصابة بالسرطان، في حين أن عوامل أخرى، مثل جانب الورم وTSH، تظهر ارتباطات أضعف. يوضح نموذج بروبيت توافقًا متوسطًا قويًا مع البيانات، وقد تم جمع البيانات من برج الصحة الذكية في السليمانية، والتي تشمل 100 حالة من السرطان واضطرابات الغدة الدرقية.

تم استخدام نموذج بروبيت للتحقيق في العلاقة بين هذه العلامات وحالات الإصابة بالسرطان لتقدير المخاطر. تشير النتائج إلى أن بعض العوامل، مثل جانب الورم ومستويات هرمون تحفز الغدة الدرقية (TSH)، كانت أقل أهمية، ولكن عوامل أخرى، بما في ذلك الجنس، غزو العقد الليمفاوية، ومستويات ATPO قبل الجراحة، أظهرت ارتباطًا قويًا بزيادة خطر الإصابة بالسرطان.

يتم استخدام نموذج الانحدار Probit لتقييم تأثير هذه المتغيرات في التنبؤ بخطر الإصابة بالسرطان والتعامل مع الطبيعة الثنائية لحالات الإصابة بالسرطان (سواء كان السرطان موجودًا أم لا). يتم تقييم ملاءمة النموذج وقوته التحليلية باستخدام اختبارات إحصائية مثل اختبار مربع كاي واختبار نسبة الاحتمالية.

وخلصت الدراسة إلى أن نمط الحياة والعوامل الديموغرافية، مثل مستويات ATPO قبل الجراحة وغزو العقد الليمفاوية، تلعب دورًا مهمًا في التأثير على خطر الإصابة بالسرطان، في حين أن عوامل أخرى مثل جانب الورم ومستويات TSH لها تأثير أقل. أظهر نموذج الانحدار Probit قدرة قوية على التنبؤ بحدوث السرطان، موضحًا جزءًا كبيرًا من تقلباته. تشير هذه النتائج إلى أن التدخلات المستهدفة والاستراتيجيات الوقائية التي تركز على عوامل الخطر القابلة للتعديل يمكن أن تقلل بشكل فعال من حالات الإصابة بالسرطان.

الكلمات المفتاحية: الانحدار، تحليل (بروبيت)، سرطان الغدة الدرقية، المتغيرات الديموغرافية، نموذج الاحتمال ثنائي المتغير، ATPO.

1. Introduction:

Thyroid cancer, one of the most prevalent endocrine malignancies, has witnessed a steady increase in incidence worldwide, raising concerns about its underlying causes and risk factors. While genetic predispositions play a critical role, lifestyle and demographic factors have emerged as significant contributors to cancer etiology. Understanding these factors and their relative influence is essential for developing targeted prevention strategies and improving public health outcomes.

A multitude of factors, such as genetic predisposition, environmental exposures, lifestyle choices, and demographic characteristics, can influence the intricate and diverse disease referred to as cancer. Comprehending the interaction of these variables is crucial for developing effective preventive strategies and influencing public health policies. Although individual risk factors have garnered significant academic attention, comprehensive models that may concurrently incorporate several variables and provide an in-depth understanding of their collective influence on cancer incidence are increasingly essential. (Srisathan et al., 2023)

This study employs a probit regression model, a statistical technique commonly utilized for examining binary or categorical outcomes, to assess the impact of various lifestyle and demographic factors on cancer incidence. The probit model, unlike linear regression, is most effective in situations where the dependent variable is binary, exemplified here by the presence or absence of cancer. The likelihood of developing cancer can be evaluated in relation to the presence or absence of chronic lymphocytic thyroiditis (CLT) based on a number of variables, such as age, gender, tumor size, multifocality, vascular invasion, capsular invasion (CI), extrathyroidal extension (ETE), involvement of lymph nodes (LN), and distant metastases. utilizing the approach of probit regression.

This introduction establishes the foundation for a comprehensive analysis of the study's methodology, results, and implications, which will enhance understanding of the modifiable risk factors associated with cancer and their interplay with demographic characteristics.(Han and Lee, 2018)

This study employs a Probit regression model to investigate the impact of various lifestyle and demographic variables on thyroid cancer incidence. The Probit model, widely used in statistical analysis for binary outcomes, offers a robust framework for identifying significant risk factors and evaluating their effects. By analyzing data from diverse sources, this research aims to uncover meaningful associations and provide insights into the complex interplay of factors contributing to thyroid cancer development.(Abonazel et al., 2023)

Doi: www.doi.org/10.25130/tjaes.21.71.1.28

2. Literature Review:

A probit regression model is a methodology used in fields like engineering, economics, and medical sciences to establish relationships between a binary dependent variable and explanatory variables. It uses the maximum likelihood estimator to determine parameters, with explanatory variables considered orthogonal. Ridge regression (RR) is an alternative estimation method for the linear regression model, developed by Kennard and Hoerl(Golam Kibria & Saleh, 2012).

The bivariate probit model, a parametric model, has been extensively used in research. However, its distributional assumption is susceptible to misspecification due to its reliance on convenience rather than economic theory. The model focuses on average treatment parameters and specific structural parameters, including binary endogenous regressors. Misspecification can pose a greater challenge in estimating many treatment parameters compared to estimating individual values, as the mean treatment parameters are expressed as the difference between marginal distributions of ε. (Han and Lee, 2018)

used of probit regression in the study provides a robust statistical framework for analyzing the factors influencing open eco-innovation among SMEs in Thailand. By employing this model, the research effectively captures the binary outcomes of firms' engagement in eco-innovation practices based on various predictors related to open innovation activities.
Used from the probit regression analysis reveal that certain forms of support and participation in open innovation programs significantly affect the likelihood of SMEs adopting eco-innovation strategies. Specifically, the results indicate that SMEs with uncertain internal open innovation processes are less likely to engage in eco-innovation, highlighting the importance of clarity and strategic direction in innovation initiatives. Moreover, the analysis demonstrates that investments in research and development, as well as networking activities, play a critical role in enhancing eco-innovation capabilities. The study emphasizes the necessity for SMEs to carefully evaluate the benefits and challenges associated with open innovation to optimize their eco-innovation outcomes. (Srisathan et al., 2023).

The authors highlighted that ridge regression can significantly reduce the variance of coefficient estimates in the presence of multicollinearity among predictor variables. They emphasize that the simulations demonstrate how ridge regression can provide more reliable estimates compared to ordinary least squares regression, particularly in scenarios where predictors are highly correlated. Additionally, the paper discusses the choice of the ridge parameter and its impact on the bias-variance trade-off, suggesting that careful selection can lead to improved model performance. The authors also note the importance of understanding the underlying data structure and the conditions under which ridge regression is most beneficial. (Hoerl et al., 1975)

The paper concluded that the analysis of children's work status and the factors influencing violence against women in Turkey reveals significant demographic and social variables. In the study by Erkan, the multinomial logistic regression model effectively identified key factors such as rural/urban status, gender, age group, household size, literacy, school attendance, and the education level of the head of the household as significant predictors of children's work status. Similarly, in the context of violence against women, variables such as the education level of women, the husband's work sector, and various relational dynamics were found to be statistically significant in influencing the types of violence experienced. (Salh, A.P.D.S.M., Abdalla, H.T. and Omer, Z.M., 2021.)

The article proposed strategies for estimating multi-collinearity in PRORM regression models, addressing the challenge of estimating the non-linear biasing parameter for the probit ridge estimator. It also highlights the need for additional estimators in the Probit regression, as other one- and two-parameter estimators surpass the ridge in several models. The primary objective is to implement alternative effective biased regression estimators across various models for the Probit regression. (Abonazel et al., 2023)

Cancer is a growing health concern in Iran, ranking third among known causes of mortality after accidents and cardiovascular disease. Gliacoma is the most prevalent cancer among Iranian men, second to breast cancer in women. Colorectal cancer is the most common gastrointestinal disorder, with a high mortality rate. The World Health Organization reports approximately 87,500 new cases diagnosed annually. Esophageal cancer, the second leading cause of cancer-related fatalities globally and the fourth most common malignancy, remains a significant public health issue. Logistic regression is the most commonly used statistical method to analyze the relationship between a binary outcome and several variables. Probit regression is an alternative log-linear technique for handling categorical dependent variables. This study aimed to evaluate logit and probit models for gastrointestinal cancer, focusing on the role of esophageal cancer in Iran.(Costa-Font and Gil, 2005).

3. Statistical methodology: The probit model is given by

$$y_i^* = x_i' \beta + \varepsilon_i; \quad i = 1, 2, \ldots, n, \qquad (1)$$

where y_i^* is an unobserved or latent variable, $x_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is defined as the i^{th} row of an X matrix with dimension $n \times p$; p is the number of the explanatory variables, β is a $(p \times 1)$ vector of the regression coefficients, and ε_i is an error term that follows a normal distribution. Since the latent is unobservable in real-world datasets, the following dummy variable is examined instead.(Golam Kibria and Saleh, 2012; Khaleel and Al-Jamil, 2021)

4. Regression:

Using one or more independent or explanatory variables (designated X1, X2, etc.) as the foundation, regression analysis creates an explanatory or predictive model of a dependent or response variable (often denoted Y). When the dependent variable is a count, such the frequency of an occurrence or the total number of individuals in a particular group, Poisson regression is suitable. If several observations have really low values, it could be especially helpful. In the field of geography as well as the social and environmental sciences in general, the methodology is still not well recognized. (Othman & Badal, 2023).

5. Probit Regression:

The binary or dichotomous outcome variables are modeled using the generic probit regression model. In microeconomics, health economics (Bishai, 1996), and medical research (see Bhattacharyya, 1997), the probit regression model is frequently employed when the goal of the study is to model a latent variable, Yn, that may be specified using the regression relationship:

$$y_i^* = x_i' \beta + u_i, i = 1, 2, \ldots, n \qquad (2)$$

- x_i' is the transpose of the i-th row of the data matrix X,
- β is the vector of regression coefficients, and
- u is the error term.

where x_i is the ith row of the $n \times (p+1)$ data matrix, X with (p+1)-vector, β of regression coefficients and $u = (u_1, u_2, \ldots, u_n)'$ is the error vector with iid components having a continuous CDF, F(x) defined on R^1 with finite Fisher information(Salh et al., 2021):

$$I(f) = \int_{-\infty}^{\infty} \left[-\frac{f'(u)}{f(u)}\right]^2 f(u)\, du, \qquad (3)$$

where f is the pdf of u and $f'(u)$ is the derivative of f(u). On the other hand, the latent variable $Y^*$is unobservable. So instead of y^*, we get the following dummy variable:

$$y_i = \begin{cases} 1 \text{ if } & y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

It has the following parameter and is distributed as a Bernoulli variable:

$$\pi_i = F(x_i'\beta); \quad i = 1, 2, \dots, n \qquad (5)$$

Given a sample size of n, the likelihood function is given by

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}. \qquad (6)$$

Hence,

$$\ln L(\beta) = \sum_{i=1}^{n} y_i \log F(x_i'\beta) + \sum_{i=1}^{n} (1-y_i) \log(1 - F(x_i'\beta)) \quad (7)$$

The ML equations are utilized to estimate the success parameters β. The asymptotic characteristics of the several kinds of probit ridge regression estimators are examined in this research(Gibbons et al., 1994).

A probit model is employed by the study for its estimations. According to Han and Lee (2019), a statistical model called a probit regression is frequently used to estimate the likelihood of an occurrence, such as how SMEs would react to eco-innovation. It works especially well with binary results. The binary outcome is tested using the probit regression model, where 'o' stands for 'unlikely' and '1' for 'likely'. The binary probit regression equation is configured as follows (Nazeera Sedeeq Barznji et al., 2024)

$$\Pr(Y = 1 \mid X) = \beta_0 + \beta_1 X \qquad (8)$$

Where:

X: is the explanatory variable.

β_0: is the intercept parameter of probit regression model.

β_1: is the slope parameter of probit regression model.

5-1. A Random-Effects Probit Regression Model: To discusses the random-effects probit regression model, introduced by Gibbons and Bock in 1987, and its generic parameter estimation approach. It also discusses empirical Bayes estimates of person- or cluster-specific effects and standard errors, allowing trends to be assessed at both group and individual levels.

The paper presents a single random effect model suitable for clustered or longitudinal research designs and a model with two random effects for longitudinal data analysis. (Srisathan et al., 2023)(Alhemoud, 2011; Mawlood and Qader, 2021; Obed and Saleh, 2023)

6. **Chronic lymphatics thyroid (CLT):** Papillary thyroid tumors (PTCs) are the most common type of thyroid cancer, accounting for 30% of all thyroid cancers. The relationship between PTC and chronic lymphocytic thyroiditis (CLT) is debated, with estimates ranging from 0.5 to 38%. The pathogenic mechanism for PTC and CLT development is still unclear, with theories including immunological reactions, preexisting CLT, or an imbalance between apoptotic and antiapoptotic pathways. Combining PTC with CLT has been linked to better prognosis, decreased recurrence rates, and less severe clinical presentation. However, some researchers have shown no benefit from cohabitation of CLT. This study aimed to determine clinicopathological factors associated with PTC prognosis in patients with or without concurrent CLT(Mawlood and Qader, 2021; Abonazel et al., 2023).

7. **Application Part**

When analyzing binomial response variables, probit analysis is a kind of regression that is employed. The functional link between an ordinal dependent variable and a collection of independent variables is ascertained by ordered probit regression. A specific type of regression model for binomial response variables is called Peobit Analysis. Regression analysis, as you may recall, involves fitting a line to your data in order to compare the relationship between the independent variable (X) and the response variable (Y).

$$Y = a + bX + e \qquad \textbf{(9)}$$

Where:

a= y-intercept

b= the slope of the line

X: is the explanatory variable

Y: is the response variable

e= error term

Also remember that a binomial response variable refers to a response variable with only two outcomes.

The study "Estimating the Impact of Lifestyle and Demographic Factors on Thyroid Cancer Incidence Using a Probit Regression Model"

Doi: www.doi.org/10.25130/tjaes.21.71.1.28

examines how different lifestyle and demographic factors may affect cancer incidence using statistical techniques.

7-1. Objective of the Study: The study's main objective is to evaluate the effects of various lifestyle decisions and demographic variables (such as age, gender, and medical history) on the risk of acquiring cancer. The research employs a Probit Regression Model, a statistical technique frequently employed in situations where the dependent variable is binary (that is, where the result is either "yes" or "no," like in the case of cancer or not).

Similar to logistic regression, probit regression models the association between the dependent variable (Thyroid cancer incidence) and the independent variables (age, gender, etc.) using a probit function, the model fits a line to the data that shows the likelihood that an event in this case, cancer will occur. The model's general form is:

$$Y = a + bX + e \qquad (10)$$

where:

- Y is the dependent variable (cancer incidence).
- X represents the independent variables (factors like age, gender, pre-op ATPO, etc.).
- a is the intercept.
- b represents the slope (impact of each factor).
- e is the error term.

7-2. Statistical Tests

How effectively the independent variables explain the variability in the dependent variable (cancer incidence) is indicated by the Pseudo R2 value.

7-3. Interpretation of Results

The findings are explained to highlight the variables that have a major impact on the incidence of cancer. For instance: o The statistical significance of age and gender suggests that they have a quantifiable effect on cancer risk. The lack of significance for several factors, such as tumour laterality and TSH (thyroid stimulating hormone), suggests that they may not have a substantial impact on the incidence of cancer in this model.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_7 X_7 + \beta_9 X_9 + \beta_{12} X_{12} + e_i \qquad (11)$$

Doi: www.doi.org/10.25130/tjaes.21.71.1.28

Table (1): Estimating Parameters of Probit Model

Parameter Estimates							
	Parameter	**Estimate**	**Std. Error**	**Z**	**Sig.**	**95% Confidence Interval**	
						Lower Bound	**Upper Bound**
Probit[a]	Age	-.022	.023	-.975	.030	-.066	.022
	Gender	-.301	.751	-.401	.008	-1.773	1.170
	pre op ATPO	.005	.001	4.386	.000	.003	.007
	PRE OP (TIRADS)	-.033	.264	-.124	.002	-.551	.485
	Tumer laterality	.034	.076	.453	.050	-.114	.183
	Lymph node invasion	.912	.383	2.383	.017	.162	1.661
	TSH	.032	.100	.319	.750	-.165	.228
a. Probit Model: Probit(P) = Intercept + Bx							

And a summary of the interpretation of each parameter:

A. **Age:** The age estimate is -0.022, with a Z-value of -0.975 and a standard error of 0.023. The statistical significance of age is indicated by the p-value of 0.030, which is less than 0.05. The range of the 95% confidence interval is -0.066 to 0.022.

B. **Gender**: The Z-value is -0.401, the gender estimate is -0.301, and the standard error is 0.751. With a p-value of 0.008, statistical significance is shown. The range of the confidence interval is -1.773 to 1.170.

C. **3. Pre-op ATPO:** The Z-value is 4.386, the estimate is 0.005, and the standard error is 0.001. With a p-value of 0.000, the data is statistically significant. The range of the confidence interval is 0.007–0.003.

D. **Pre-operation (TIRADS):** The estimate has a Z-value of -0.124 and a standard error of 0.264. It is -0.033. There is statistical significance as indicated by the p-value of 0.002. The range of the confidence interval is -0.551 to 0.485.

E. **Tumour Laterality:** Z-value is 0.453, standard error is 0.076, and the estimate is 0.034. With a p-value of 0.050, the result is only marginally significant. The range of the confidence interval is -0.114 to 0.183.

F. **Lymph Node Invasion:** Z-value is 2.383, standard error is 0.383, and the estimate is 0.912. With a p-value of 0.017, statistical significance is indicated. The range of the confidence interval is 0.162 to 1.661.

G. **TSH:** The estimate is 0.032, with a Z-value of 0.319 and a standard error of 0.100. A p-value of 0.750 suggests that there is no statistical significance. The significance values and confidence intervals suggest that some variables (like Pre-op ATPO and Lymph Node Invasion) are strongly associated with the outcome, while others (like TSH and Tumour Laterality) might not be.

Table (2): Covariance's and Correlations of Parameter Estimates of probit analysis

		Age	**Gender**	**pre op ATPO**	**Pre op (TIRADS)**	**Tumor laterality**	**Lymph node invasion**	**TSH**
Probit	Age	.001	-.087	.198	-.112	.163	-.167	-.300
	gender	-.001	.564	.111	-.020	-.016	.048	.102
	pre op ATPO	.000	.000	.000	-.389	.243	.028	-.246
	Pre op (TIRADS)	-.001	-.004	.000	.070	-.075	-.078	-.069
	Tumor laterality	.000	-.001	.000	-.002	.006	-.111	-.281
	Lymph node invasion	-.001	.014	.000	-.008	-.003	.146	-.062
	TSH	-.001	.008	.000	-.002	-.002	-.002	.010
Covariances (below) and Correlations (above).								

Table 2 presents covariances below the diagonal and correlations above the diagonal. Age demonstrates associations ranging from a negative correlation of -0.300 with TSH to a positive correlation of 0.198 with pre-op ATPO. The maximum covariance (age with itself) is only 0.001, indicating a very small value. Gender and age show a somewhat positive association (0.564) and a moderately positive correlation (0.102) with TSH. Most other correlations and covariances are negligible or very small.

The table effectively represents the correlations among the variables, accurately reflecting the data provided. The weakest correlations involving age are with pre-op ATPO (0.198, positive) and TSH (-0.300, negative). Age-related covariances are minimal, with the highest value being 0.001 for its covariance with itself. Gender and age display a somewhat positive association (0.564) and a moderately positive correlation (0.102) with TSH. While there are weaker or negative associations elsewhere, the strongest positive association is observed between pre-op ATPO and TSH (0.028). Conversely, a significant negative correlation (-0.389) is noted between pre-op ATPO and pre-op TIRADS.

The least strong correlations, or little to no linear association, are seen between tumor laterality and the other factors. Additionally, the covariance's would be small due to the poor correlations.

There is a rather significant positive connection between this variable and TSH (0.146). A noteworthy association is shown by the moderate covariance value of the correlation with TSH. TSH has the strongest positive correlation (0.028). With PRE OP (TIRADS), it has a somewhat negative association (-0.389). With the exception of a tiny negative correlation (-0.075) with tumor laterality, the correlations with the remaining factors are largely minor. With values almost equal to zero, it has the lowest correlations. With TSH, it exhibits a rather strong positive connection (0.146). There are negligible associations with the other variables; lymph node invasion has the most positive correlation (0.146).

Table (3): Chi-Square Tests of probit analysis

Chi-Square Tests				
		Chi-Square	**df[a]**	**Sig.**
Probit	Pearson Goodness-of-Fit Test	21.997	93	0.0001
	Likelihood ratio test	125.79		
	Pseudo R^2	0.5025		
	Log- Likelihood	-141.27		
	Number of Observation	100		
a. Statistics based on individual cases differ from statistics based on aggregated cases.				

This table above shows that the model is significantly fitted, as indicated by the value of (Chi-Square = 21.997) at significant alpha = 5%) and p-value 0.0001 in the above table. Only when compared to another

Doi: www.doi.org/10.25130/tjaes.21.71.1.28

pseudo-R square of the same type on the same data and forecasting the same outcome can a faux R square have significance. In this case, the model that best predicts the outcome is shown by a greater pseudo-R squared, an overview of the Chi-Square test findings and other statistics pertaining to a Probit regression model. Here is how the important statistics are interpreted:

Pearson Goodness-of-Fit Test:

Chi-Square: 21.997

Degrees of Freedom (df): 93

Significance (Sig.): 0.0001

This shows that the p-value for the Pearson Goodness-of-Fit test is 0.0001 and the chi-square value is 21.997 with 93 degrees of freedom. The null hypothesis that the model does not fits the data well is rejected when the p-value of 0.0001 indicates that the model matched the data well since the significance is higher than the usual threshold of 0.05.

A Pseudo R^2 of 0.5025 suggests that the model accounts for approximately 50.25% of the variation in the dependent variable, representing a reasonably strong fit and indicating the model’s robustness.

In the context of maximum likelihood estimation, the log-likelihood value provides insight into the probability of the observed data given the model. A log-likelihood value of -141.27 suggests the likelihood of the data under the model, where lower (more negative) values generally indicate a better model fit.

Over-all, a higher value indicates a better fit compared to a simpler model, such as one without predictors. The model probit regression in goodness of fit was assessed using the Likelihood Ratio Test, which yielded a result of 125.79. This value reflects the extent to which the independent variables explain the variance in the dependent variable.

8. Conclusion

The statistical analysis indicates that the study's result is likely to reveal a strong correlation between some lifestyle and demographic factors and the incidence of thyroid cancer, whereas other elements may exert a lesser influence. The Probit model aids in finding significant cancer risk factors and provides a statistical framework for evaluating these relationships.

The study used a robust probit regression model to identify and evaluate key lifestyle and demographic factors that affect thyroid cancer

Doi: www.doi.org/10.25130/tjaes.21.71.1.28

incidence. Although the model demonstrates moderate to high accuracy through pseudo-R2 measurements, But diagnostic testing revealed potential compliance issues. This indicates the need for further refinement. Gender-specific preoperative ATPO levels and lymph node invasion were significantly positively associated with cancer risk. This can be seen from the statistically significant estimates and confidence intervals.

The pseudo-R2 indicated that the Probit model fit the data with moderate to high precision. The Pearson goodness-of-fit test and the log-likelihood score of -141.27 suggest potential issues with the model's specification, indicating that the model may accurately represent the data.

The gender estimation was conducted pre-operative ATPO. An estimate of 0.005 and a p-value of 0.000 suggest a robust positive correlation for this variable, signifying a meaningful impact. Lymph Node Invasion variable exhibited a p-value of 0.017 and a significant positive estimate of 0.912. The study generated 95% confidence intervals to explain the range within which the genuine population parameters are expected to be a feature of. The pre-operative ATPO confidence interval demonstrated a substantial association, spanning from 0.003 to 0.007. The research highlighted the importance of p-values in evaluating the significance of each variable, with values under 0.05 indicating statistical significance for most of the components analyzed.

The study highlights a significant relationship between factors like preoperative ATPO and lymph node invasion with cancer likelihood, emphasizing their potential predictive value. However, the Probit model's limitations, such as the low log-likelihood score, suggest that its predictive accuracy could be enhanced by incorporating more explanatory variables or exploring alternative modeling approaches.

The difference in confidence intervals further underscores the differing levels of influence among the factors studied, indicating areas where additional analysis may clarify their roles. The recommendation for further research is well-founded, aiming to refine causal interpretations and improve early detection strategies by leveraging the identified factors more effectively.

References:

1. Abonazel, M.R. et al. (2023) 'New estimators for the probit regression model with multicollinearity', Scientific African, 19. Available at: https://doi.org/10.1016/j.sciaf.2023.e01565 .
2. Alhemoud, A.M. (2011) An ordered probit regression model for estimating the effects of demographic factors on rice consumption in Saudi Arabia, Int. J. Business and Globalisation. Available at: http://www.oryza.com/mideast/Saudiarabia/saudiarabiatables2.shtm1.
3. Babli, S., Payne, R., Mitmaker, E., & Rivera, J. Effects of Chronic Lymphocytic Thyroiditis on the Clinicopathological Features of Papillary Thyroid Cancer. Eur Thyroid J. 2018; 7 (2): 95–101.
4. Costa-Font, J. and Gil, J. (2005) 'Obesity and the incidence of chronic diseases in Spain: A seemingly unrelated probit approach', Economics and Human Biology, 3(2 SPEC. ISS.), pp. 188–214. Available at: https://doi.org/10.1016/j.ehb.2005.05.004.
5. Do, Y. K., & Wong, K. Y. (2012). Awareness and acceptability of human papillomavirus vaccine: an application of the instrumental variables bivariate probit model. BMC public health, 12, 1-8.
6. Gibbons, R.D., Hedeker, D. and Lab, B. (1994) Application of Random-Effects Probit Regression Models, Journal of Consulting and Clinical Psychology.
7. Golam Kibria, B.M. and Saleh, A.K.M.E. (2012) 'Improving the estimators of the parameters of a probit regression model: A ridge regression approach', Journal of Statistical Planning and Inference, 142(6), pp. 1421–1435. Available at: https://doi.org/10.1016/j.jspi.2011.12.023 .
8. Han, S. and Lee, S. (2018) 'Estimation in a Generalization of Bivariate Probit Models with Dummy Endogenous Regressors'. Available at: http://arxiv.org/abs/1808.05792.
9. Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression: some simulations. Communications in Statistics-Theory and Methods, 4(2), 105-123.
10. Khaleel, Z.J. and Al-Jamil, S.K. (2021) 'The impact of Some Public Finance variables on the Gross National Savings in Iraq', Tikrit Journal of Administrative and Economic Sciences, 17(54, 3), pp. 430–443. Available at: https://doi.org/10.25130/tjaes.17.54.3.27.
11. Kibria, B. G., & Saleh, A. M. E. (2012). Improving the estimators of the parameters of a probit regression model: A ridge regression approach. Journal of Statistical Planning and Inference, 142(6), 1421-1435.
12. Mawlood, K.I. and Qader, H.M. (2021) 'Estimating Hazard Function and Survival Analysis of Tuberculosis Patients in Erbil city', Tikrit Journal of Administrative and Economic Sciences, 17(54, 3), pp. 473–492. Available at: https://doi.org/10.25130/tjaes.17.54.3.30 .
13. Nazeera Sedeeq Barznji et al. (2024) 'Utilizing Cox Regression Analysis and Bootstrapping test for Indicating the Risk of Smoking on Health in Kurdistan', Tikrit

Journal of Administrative and Economic Sciences, 20(65, part 1), pp. 331–347. Available at: https://doi.org/10.25130/tjaes.20.65.1.19 .

14. Othman, S. and Badal, M.A. (2023) 'Using Multilevel Poisson Regression with Wavelets Filters', University of Kirkuk Journal for Administrative and Economic Science, 13(4), pp. 179–192.
15. Pourhoseingholi, A., Pourhoseingholi, M. A., Vahedi, M., Safaee, A., Moghimi-Dehkordi, B., Ghafarnejad, F., & Zali, M. R. (2008). Relation between demographic factors and type of gastrointestinal cancer using probit and logit regression. Asian Pac J Cancer Prev, 9(4), 753-755.
16. Salh, S.M., Abdalla, H.T. and Omer, Z.M. (2021) 'Using Multinomial Logistic Regression model to study factors that affect chest pain', Tikrit Journal of Administrative and Economic Sciences, 17(53, 2), pp. 534–555. Available at: https://doi.org/10.25130/tjaes.17.53.2.31 .
17. Srisathan, W.A. et al. (2023) 'Driving policy support for open eco-innovation enterprises in Thailand: A probit regression model', Journal of Open Innovation: Technology, Market, and Complexity, 9(3). Available at: https://doi.org/10.1016/j.joitmc.2023.100084 .
18. Obed, S. and Saleh, D. (2023) 'The Impact of Social Media Advertising on Customer Performance Using Logistic Regression Analysis', Qalaai Zanist Scientific Journal, 8(3). Available at: https://doi.org/10.25212/lfu.qzj.8.3.54.