نموذج كمي للتعرف على المحددات الرئيسة لمطالبات التأمين الصحي في المملكة العربية السعودية

محمد عبد المولى عثمان
قسم المالية – جامعة الملك سعود
قسم الإحصاء والتأمين – جامعة طنطا
maosman@ksu.edu.sa

عماد عبد الجليل علي إسماعيل
قسم التحليل الكمي – جامعة الملك سعود
قسم الرياضة والتأمين – جامعة القاهرة
emadali@ksu.edu.sa

الكلمات المفتاحية: التأمين الصحي، مبلغ المطالبة، النموذج اللوجستي، العوامل الديموجرافية، المملكة العربية السعودية.

ملخص البحث: تطور قطاع التأمين في المملكة العربية السعودية بدرجة كبيرة في السنوات الأخيرة، على وجه الخصوص التأمين الصحي وتأمين السيارات. ورغم هذا التطور في سوق التأمين السعودي إلا أن بعض شركات التأمين في هذا السوق حققت خسائر باهظة مما اضطرها إلى الخروج من السوق نتيجة تأثير هذه الخسائر على مراكزها المالية. وهذا البحث يستكشف المحددات الرئيسية التي تؤثر على مطالبات التأمين الصحي الفردي. وبمعنى آخر يحاول هذا البحث استكشاف ماهية العوامل الديموجرافية لحملة الوثائق التي تؤثر على مبالغ مطالبات التأمين الصحي في سوق التأمين السعودي؟ حيث إن تحليل العوامل المؤثرة على المطالبات يفيد في التحكم في الخطر وانتقاء الأخطار التي تؤمن عليها شركات التأمين، فضلاً عن مساعدتها في تطبيق إجراءات الاكتتاب الملائمة. ولتحليل محددات مطالبات التأمين الصحي تم استخدام بعض النماذج الإحصائية على وجه الخصوص النموذج الإحصائي اللوجستي لاستكشاف المحددات الأساسية لهذه المطالبات. وقد خلص هذا البحث ببعض المحددات التي يمكن لمكتتبي التأمين الاسترشاد بها سواء عند تجديد الوثائق السارية أو اكتتاب وثائق جديدة.

Aly, I. A. Medical or health insurance: Problems and solutions.

Chen, T. (2003).Risk factors of medical expense in China and statistical models. Managerial Finance, 29(5/6), 52–64.

Chen, X. X., & Zhao, Y. M. (1994). The expense controlling system of children's hospital expense insurance in Shanghai. China Hospital Management, 14(8), 24–27.

Edwards, T. C. (2003). Assuming association: Logistic regression and logit analysis. Biometry, FRAWS 6500,Fall.

Gebotys, R. J. (2000). Examples: Binary l Logistic Regression, January.

Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression. New York: John Wiley and Sons, Inc.
http://stp.lingfil.uu.se/~nivre/statmet/strombergsson.pdf.
http://www.al-sharq.com/news/details/359753.

https://pdfs.semanticscholar.org/7fa6/1d9639ff581fe6bfccdb96c5986a07b0fb57.pdf

Ismail, A. A., Hamza, A., & Al-Hudief, S. (2013a).The challenges of the development of medical insurance in Saudi Arabia. Egyptian Journal of Insurance and Actuarial Science, College of Commerce, Cairo University.

Ismail, A. A., Hamza, A., & Al-Hudief, S. (2013b). Quantitative model to finance the net cost of medical insurance forlow-income people. Egyptian Journal of Insurance and Actuarial Science, College of Commerce, Cairo University.

Ismail, E.A. (2016). Support the decision of insurance company while underwriting individual health insurance. Journal of Commerce and Finance, College of Commerce, Tanta University, EGYPT.

Khodair, A. (2012). Using logistic regression model in prediction of dichotomous economic-dependent variables. Journal of Karkook University for Economics and Business, 2(2).

Lambon, Quayefio and Nkechi S. Owoo (2017) Determinants and the impact of the National Health Insurance on neonatal mortality in Ghana ,Health Economics Review (2017) 7:34

Logistic regression. IJRRAS, 10(1).

Mohammad Ranjbar. E, Ameneh Khosravi, Mohammad Amin. B, Sima Rafiei, (2018) "Socio-economic inequalities in health services utilization: a cross-sectional study", International Journal of Health Care Quality Assurance, Vol. 31 Issue: 1, pp.69-75,

Overpeck, M. D., Jones, D. H., & Trumble, A.C. (1997). Socioeconomic and racial-ethnic factors affecting non-fatal medically attend injury rates in U.S. children. Injury Prevention,3(4), 272–276.

Peng, C. Y., Kuk, L., & Ingersoll, G.M. (2002).An introduction to logistic regression analysis and reporting. Journal of Educational Research, 96(10), 3–13.

Rush, S. (2001).Logistic regression: The standard method of analysis in medical research.

Saudi Arabian Monetary Agency (SAMA).over period 2007- 2017.

Scott E. Harrington, S. E. (2010). The health insurance reform debate. Journal of Risk and Insurance, 77(1), 5–38.

Seng, H. L., Xu, D. F., & Wang, L. X. (1992). Discussing several factors that affect public medical expense. China Hospital Management, 12(7), 32–35.

Spicer, J. (2004). Making sense of multivariate data analysis: An intuitive approach. California: Sage Publications, 123–151.

Strömbergsson, S. (2009). Binary logistic regression and its application to data from a study of children's recognition of their own recorded voice.

Walker, M.D. (1998). Discriminant Function Analysis, Lesson8.

Yusuff, H. N., Mohamad, N., & Yahaya, A. S. (2012). Breast cancer analysis using

Zhang, H. N. (1996).Overviews about the reason and strategy for the high claim ratein hospital insurance. Insurance Research, (3), 51–53.
(https://www.mubasher.info/news/3072903)

Consequently, the probabilities of classifying the claim (according the marital status2 variable) will be as follows:

$$Y = \frac{ODDS}{1+ODDS} = \frac{3.732}{4.732} = 78.9\%$$

Therefore, the model predicts that 78.9% of claims had rated as low.

Further,

$$Y = \frac{ODDS}{1+ODDS} = \frac{86.747}{1+86.747} = 98.9\%$$

Thus, the model predicts that 98.9%of claims had rated as high. Moreover, Table (12) indicates the value of Exp(B) for the marital status 2 variable is 25.254. This means the classification of claims from low to high is 25.254 times greater for those who are not single, as opposed to single, at confidence intervals between 3.042 and 177.762,as non-single persons have a higher usage rate of hospitals than those who are single.

We use the odds ratio Exp(B) to rank significant predictor variables according to their importance, as shown below:

Table (13) Predictor Variables According to their Importance

Rank	1	2	3	4	5
Variable	Marital2	Marital1	Gender	Nationality	Age

In conclusion, we note the previous analysis had derived an optimal logistic model to determine the key determinants for health insurance claims in the Kingdom of Saudi Arabia. This model is

Log (odds) = 1.317 – 0.130 age +1.33 gender +0.300 nationality + 2.764 marital1 + marital2 3.146

This model considers an effective tool for selecting potential insureds in the Saudi health insurance market. Moreover, it is a guide for the underwriter while renewing current policies at a moderate rate.

5.5 Conclusion

High ratios of individual health insurance claims to premiums in the Saudi insurance market have resulted in some insurance companies in that market incurring excessive losses and hence exiting the market. Thus, the researchers have attempted to study the demographic factors that affect health insurance claims in the Kingdom of Saudi Arabia by using data available from large insurance company in the Saudi insurance market. The researchers used a logistic regression analysis to reveal the important demographic factors of the insured that affect claims. The study revealed five demographic factors: age, gender, nationality, and two types of marital status factors. The classification of individual health insurance claims using these demographic factors was successful (90.8%). As demographic factors differ among policyholders, this research's results will be a guide for underwriters in making crucial and successful decisions while rating health insurance policies. Moreover, this study provides insights specifically for underwriters in insurance companies, in particular, when they renew the current health insurance policies or pricing the new health insurance policies. This study's findings are important and relevant to the different lines in the insurance industry, and significantly contribute to our society in general and insurance companies in particular.

5.6 Recommendations for further research

As health insurance in the Kingdom of Saudi Arabia is compulsory, all employers obligated by law to buy health insurance policies for their workers. Hence, the number of insured persons has increased over time, and both health care providers and insurance companies need more research. The authors recommend further research using other demographic factors of the insured, such as occupation, area of residence, income, and education, as these factors are currently unavailable and will reveal the diversity in degrees of risk among the insured. These factors can be utilized to calculate appropriate premiums for each insured and realize the equity among them.

ACKNOWLEDGEMENTS

The researchers are grateful to the Research Center at the College of Business Administration (CBA) and the Deanship of Scientific Research at King Saud University Riyadh for their financial support of this research.

References

Achia, T. N., Wangombe, A., & Khadioli, N. (2010).A logistic regression model to identify key determinants of poverty using demographic and health survey data. European data. European Journal of Social Science, 13(1).

Al Farhood, S. H. (2014). The use of logistic regression in studying the factors influencing the performance of stocks (An empirical study on the stock exchange). Journal of Al Azhar University Gaza (Natural Science), 16, 47–68.

Allison, P. (2013). Why I don't trust the Hosmer-Lemeshow test for logistic regression. Statistical Horizons.http://statisticalhorizons.com/hosmer-lemeshow

Thus, the model predicts that 76.6% of claims had rated as high.

Moreover, Table (12) notes that the value of Exp(B) for the variable age is (0.878 <1). Therefore, the classification of claimsfrom low to high will decrease by 0.130 at a confidence interval between 0.86 and 0.896.When the insured's age increases by one year, in other words, hisor her claim classification, from low to high, increases by 0.878 on average.

- **The second predictor variable: Gender**

If all other independent predictor variables are constant, the odds of the variable gender will equal

$$\text{ODDS} = \mathbf{e}^{\mathbf{1.317+1.339(0)}} = 3.732$$

for men (male =0), and

$$\text{ODDS} = \mathbf{e}^{\mathbf{1.317+1.339(1)}} = 14.239$$

For women (female =1).

Consequently, the probabilities of classifying the claim (according to the gender variable) will be as follows:

$$\text{Y} = \frac{ODDS}{1+ODDS} = \frac{3.732}{4.732} = 78.9\%.$$

Thus, the model predicts that 78.9% of claims had rated as low.

Further,

$$\text{Y} = \frac{ODDS}{1+ODDS} = \frac{14.239}{1+14.239} = 93.4\%.$$

Thus, the model predicts that 93.4% of claims had rated as high.

Moreover, Table (12) reveals the value of Exp (B) for the gender variable as 3.815, or the classification of claims from low to high is 3.815 times greater for females than males, at confidence intervals between 2.274 and 6.402. This may be because females have higher medical service utilization rates than males.

- **The third predictor variable: Nationality**

If all other independent predictor variables are constant, the odds of the nationality variable will equal

$$\text{ODDS} = \mathbf{e}^{\mathbf{1.317+.300(0)}} = 3.732$$

If non-Saudi =0, and the ODDS = $\mathbf{e}^{\mathbf{1.317+0.300(1)}}$ = 5.038if Saudi =1.

Consequently, the probabilities of classifying the claim (according the nationality variable) will be as follows:

$$\text{Y} = \frac{ODDS}{1+ODDS} = \frac{3.732}{4.732} = 78.9\%.$$

Thus, the model predicts that 78.9% of claims had rated as low.

Further,

$$\text{Y} = \frac{ODDS}{1+ODDS} = \frac{5.038}{1+5.038} = 83.4\%$$

Thus, the model predicts that 83.4% of claims had rated as high.

Moreover, Table (12) displays the value of Exp(B) for the gender variable as 1.350. Therefore, the claim classification from low to high claim, at confidence intervals between 0.955 and 1.904, is 1.350 times for non-Saudi greaterthan Saudi. This may be because hospitals experience higher frequency rates of non-Saudis thanSaudi.

- **The fourth predictor variable: Marital Status 1**

If all other independent predictor variables are constant, the odds of the marital status1variable will equal

$$\text{ODDS} = \mathbf{e}^{\mathbf{1.317+2.764(0)}} = 3.732$$

If non-married = 0, and

$$\text{ODDS} = \mathbf{e}^{\mathbf{1.317+2.764\,(1)}} = 59.205$$

If married =1.

Consequently, the probabilities of classifying the claim (according the marital status1 variable) will be as follows:

$$\text{Y} = \frac{ODDS}{1+ODDS} = \frac{3.732}{4.732} = 78.9\%.$$

Thus, the model predicts that 78.9% of claims had rated as low.

Further,

$$\text{Y} = \frac{ODDS}{1+ODDS} = \frac{59.205}{1+59.205} = 98.3\%.$$

Thus, the model predicts that 98.3%of claims had rated as high.

Moreover, Table (12) indicates the value of Exp(B) for the marital status1 variable is 15.856. Therefore, the classification of claims from low to high is 15.856 times greater for non-marrieds as opposed to those who are married, at confidence intervals between 2.097 and 119.862. This may be because married persons use hospitals more frequently than those who are unmarried.

- **The fifth predictor variable: Marital Status 2**

If all other independent predictor variables are constant, the odds of the marital status2 variable will equal

$$\text{ODDS} = \mathbf{e}^{\mathbf{1.317+3.146(0)}} = 3.732$$

If single =0, and

$$\text{ODDS} = \mathbf{e}^{\mathbf{1.317+3.146\,(1)}} = 86.747$$

If non-single =1.

important demographic factors affecting health insurance claims, and will guide underwriters in the Saudi insurance market in making optimal decisions concerning the claim before renewing individuals' health insurance policies.

Table (11) Classification Tablea

Observed			Predicted		
			Claim		Percentage Correct
			0	1	
Step 1	Claim	0	1,381	8	99.4
		1	144	125	46.5
	Overall Percentage				90.8

a. The cut value is .500

5.4.3 Statistical tests of predictor (independent) variables

After testing the logistic model's goodness of fit, the statistical significance of individual predictors (demographic factors) is tested. Moreover, we will recognize each predictor variable's ability to predict. Further, we will know the answer to the question, "What is the priority of each predictor variable in predicting the dependent variable (claim amount). Table (12) illustrates the contributions of each independent (predictor) variable to the model, and its significance using Wald's test (as per Wald, the column is used to determine each predictor variable's statistical significance).

Table (12) Variables in the Equation

Variable		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Age	-.130	.011	152.570	1	.000	.878	.860	.896
	Gender	1.339	.264	25.705	1	.000	3.815	2.274	6.402
	Nationality	.300	.177	2.892	1	.089	1.350	.955	1.909
	Residence	-.240	.243	.980	1	.322	.786	.489	1.266
	Employment	-.161	.225	.511	1	.475	.851	.547	1.324
	Marital Status 1	2.764	1.032	7.170	1	.007	15.856	2.097	119.862
	Marital Status 2	3.146	1.038	9.193	1	.002	23.254	3.042	177.762
	Constant	1.317	1.187	1.230	1	.267	3.732		

a. Variable(s) entered in Step 1: age, gender, nationality, residence, employment status, and the two marital status variables.

By looking at Table (12), we note the following:

1. The age, gender, and marital status 1 and 2 variables are significant, with a *p*-value of less than0.05, and the nationality variable is significant, with a *p*-value of less than 0.10 and a 95% degree of confidence. Therefore, these variables are important in identifying the type of claim (high or low), and they add significance to the model and prediction.
2. The residence and employment variables are not significant, and are not important in identifying the type of claim (high or low).
3. Moreover, the odds ratio Exp(B) for each significant predictor variable may interpreted as follows:

- **The first predictor variable:Age**

We use the odds prediction equation to predict the probability of classifying a claim based on a one-unit change in the predictor variable age when all other independent predictor variables are kept constant, or:

$$\mathbf{ODDS = e^{a+bx}}$$

If the age is greater than the average age (51.71) = 1 (see Appendix 2) and less than the average age (51.71) = 0, then:

$$\text{ODDS} = \mathbf{e^{1.317\text{-}0.13(0)}} = 3.732$$

If the age is less than the average, then:

$$\text{ODDS} = \mathbf{e^{1.317\text{-}0.13(1)}} = 3.277$$

for the age that is greater than the average. Consequently, the probabilities of classifying the claim (according to the age variable) will be as follows:

$$Y = \frac{ODDS}{1+ODDS} = \frac{3.732}{4.732} = 78.9\%.=$$

Thus, the model predicts that 78.9% of claims had rated as low. Further,

$$Y = \frac{ODDS}{1+ODDS} = \frac{3.277}{1+3.277} = 76.6\%.$$

value of 436.904 for a df of 7, and α = (p-value = 0.000 < 0.05). Thus, the predictor variables are important and effectively determine the claims as high or low.

Table (7) Test of the Model's Significance (Omnibus Tests of Model Coefficients)

		Chi-square	df	Sig.
Step 1	Step	436.904	7	.000
	Block	436.904	7	.000
	Model	436.904	7	.000

Moreover, Table (8) contains two values: the first value for Cox and Snell's R-Square equals 0.232; the second for Nagelkerke's R-Square equals 0.394. These values interpret approximately 23.2% and 39.4% of the variation in claim amount (high-low), respectively. Although these values are proportionally low, they are accepted in these models, although Spicer (2004, p. 129) still advises caution regarding these "pseudo statistics," as they are "best treated with caution if not actually avoided."

Table (8) Model Summary

Step	-2 Log likelihood	Cox & Snell's R-Square	Nagelkerke's R-Square
1	1,033.317[a]	.232	.394

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .01

5.4.1.3 Hosmer and Lemeshow's test

In spite of its value, the chi-square is significant at *p*-value = 0.000, as Table (9) illustrates. We cannot say the model does not fit the data where the *p*-value is less than 0.05 because Hosmer and Lemeshow's test may be unsatisfactory (Allison 2013). Paul Alison reveals that question in his 2013 work, "Why I Don't Trust the Hosmer-Lemeshow Test for Logistic Regression." Further, Spicer (2004) has mentioned that many logistic regression analysis reports do not include information regarding the chi-square, but instead focus on the odds for each independent variable and their significance. However, the claims' observed and expected values approximate each other, as Table (10) notes, which indicates mean compatibility between the model and data.

Table (9) Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	54.667	8	0.000

Table (10) Contingency Table for Hosmer and Lemeshow's Test

		Claim = 0		Claim = 1		Total
		Observed	Expected	Observed	Expected	
Step 1	1	158	160.394	6	3.606	164
	2	144	150.225	12	5.775	156
	3	143	153.496	19	8.504	162
	4	155	153.914	10	11.086	165
	5	141	147.875	20	13.125	161
	6	159	149.579	7	16.421	166
	7	149	146.727	18	20.273	167
	8	137	134.699	24	26.301	161
	9	141	128.507	24	36.493	165
	10	62	63.584	129	127.416	191

5.4.2 Efficiency of the classification table

It is necessary to assess the effectiveness of the predictive classifications of claims against their actual classifications to judge the logistic model's efficiency. Table (11) presents the results from the classification of claims using the logistic model. An observation of Table (11) reveals that the overall success rate of the model's classification, including the predictor variables, is 90.8 %. This percentage considers high, while the percentage of error is 9.2%. However, the correct percentage of classification for high claims (46.5%) is low. Thus, approximately 53.5% of high claim amounts had rated as low (error percentage). In contrast, the error percentage of classification of low claims (0.6%) is low. Therefore, the logistic model will explore the

variables to follow a normal distribution. Second, it does not require a linear relationship between the independent and dependent variables.

It is necessary to explore the multicollinearity problem among the independent variables before analyzing data using the logistic model. An observation of these variables' correlation matrix, as indicated in Appendix (1), revealed that this problem does not exist.

Moreover, the researchers note in Table (4) that it is also necessary to construct a cross-tabulation between the dependent variable (claim amount) and the different independent variables (demographic variables) to understand the significant relationships between them.

Table (4) Values of Pearson's χ^2Statistic on Cross-Classifying Demographic Factors with Claim Amount

Explanatory Variable	χ^2 Value	df	*p*-value
Age	773.341	46	<0.00
Gender	106.384	1	<0.00
Nationality	43.195	1	<0.00
Residence	21.712	1	<0.00
Employment	6.880	1	<0.009
Marital Status 1	35.372	1	<0.00
Marital Status 2	4.401	1	<0.036

Table (4) reveals a significant relationship between the claim amount and demographic factors (the predictor variables, or the insured's age, gender, nationality, employment status, and marital status).

5.4.1 Logistic model results

We adhere to the following observations to apply the previous methodology and evaluate the logistic model's results:

5.4.1.1 Results of the logistic model without the predictor variables

Table (5) indicates the iterations of the logistic model that only contains a constant, and without the predictor variables. The maximum likelihood estimate is 1,470.221, which compare if the predictor variables are included, to recognize their effect on the dependent variable; see Table (5). Moreover, the coefficient (B0) is -1.642 and Wald's statistic is 607.323, with a significance level of 0.00, which is less than $\alpha = 0.05$. Moreover, the odds ratio Exp (B) is 0.194; therefore, we can derive the significance of the constant without predictor variables.

Table (5) Iteration of Logistic Model and Maximum Likelihood Estimates without the Predictor Variables

Iteration	-2 Log likelihood	Coefficients (B_0)	S.E	Wald	df	Sig	Exp(B)
1	1,490.518	-1.351					
2	1,470.358	-1.617					
3	1,470.221	-1.641					
4	1,470.221	-1.642	.067	607.323	1	0.000	0.194

5.4.1.2 Results of Logistic Model with the predictor variables

Using SPSS and the Enter method, we calculated Table (6), as follows:

Table (6) Iteration of Logistic Model and Maximum Likelihood Estimates with Predictor Variables[a,b,c,d]

Iteration		-2 Log Likelihood	Coefficients							
			Constant	Age	Gender	Nationality	Residence	Employment	Marital1	Marital2
Step 1	1	1,145.085	1.139	-.074	.704	.178	-.025	-.089	.925	1.193
	2	1,043.277	1.784	-.112	1.102	.266	-.151	-.134	1.586	1.919
	3	1,033.764	1.741	-.128	1.302	.297	-.228	-.157	2.237	2.607
	4	1,033.330	1.428	-.130	1.338	.300	-.240	-.161	2.649	3.032
	5	1,033.317	1.323	-.130	1.339	.300	-.240	-.161	2.758	3.141
	6	1,033.317	1.317	-.130	1.339	.300	-.240	-.161	2.764	3.146
	7	1,033.317	1.317	-.130	1.339	.300	-.240	-.161	2.764	3.146

a. Method: Enter; b. Constant is included in the model; c. Initial -2 Log Likelihood: 1,470.221

Table (6) reveals that the iterations of the logistic model with the predictor variables stops at step 7, with a maximum likelihood estimate of 1033.317; this value is less than its counterpart in Table (4). Consequently, we can calculate the goodness of fit for the model with the predictor variables, and note that a relationship exists between the claim amount as a dependent variable and the predictor variables.

Moreover, the latter has statistical significance in discrimination the claim amounts as high or low due to the significance of the logistic model, as Table (7) illustrates by observing the value of $\chi 2$. The latter has a

5.3.1 Estimating the model parameters

The logistic model's parameter scan estimate by the maximum likelihood method, as follows:

Given that the set of data observations are (x_i, y_i) and the likelihood function is $\pi(xi)$, where $y_i = 1$, and $1- \pi(x_i)$, where $y_i = 0$, and given that the following equation's result provides the contribution to the likelihood function for the observation (x_i, y_i) as $\zeta(x_i)$:

$$\zeta(x_i) = \pi(x_i)y_i\,[1- \pi(xi)\,1-y_i]. \quad (5)$$

This equation accounts for only one set of observations. As these observations are assumed independent of each other, we can multiply their likelihood contributions to obtain the complete likelihood function, the result of which is provided in Equation (6):

$$\iota(B) = \prod_{i=1}^{n} \zeta(x_i) \quad (6)$$

Where B is the collection of parameters $\beta_0, \beta_1, \beta_2 \cdots \beta_j$ and $\iota(B)$ is the likelihood function of B.

The maximum likelihood estimates can obtain by calculating B, which maximizes ι(B). We simplify this by calculating the logarithm of Equation (6) before finding the value Which maximizes the likelihood function. We obtain the following equation by calculating the logarithm of Equation (5):

$$L(B) = ln\, i(B) = \sum_{i=1}^{n} y_i\, \ln[\pi(x_i)] + (1-y_i)\, \ln[1-\pi(x_i)] \quad (7)$$

We calculate the differentiation of Equation (7) relative to $\beta_0, \beta_1, \beta_2 \cdots \beta_j$ and set the resulting derivatives equal to zero to determine the value of B that maximizes L (B). The resulting equations are called "likelihood equations," and j + 1 such equations exist; for example, note the following Equation (8) for intercept β_0:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \quad (8)$$

and the following Equation (9) for the predictor variables $\beta_0, \beta_1, \beta_2 \cdots \beta_j$:

The likelihood equations are solved using the maximum likelihood estimate B`, which is obtained using the SPSS software suite, version 21, to perform a logistic regression analysis of the data and calculate the maximum likelihood estimate, as illustrated in the following Section 5.4.

5.3.2 The logistic regression analysis

The researchers will evaluate the logistic regression model by performing many analyses, as Peng et al. (2002) recommended a data analysis of data should include the following:

5.3.2.1 An overall evaluation of the model

The overall evaluation demonstrates if the logistic regression model fits the data, and will perform twice: the logistic model will apply with and without the predictor variables. Does knowledge of the independent variables, in other words, improve the ability to predict the value of the dependent variable? This will specifically examine using likelihood ratio tests, with *p*-values less than 0.05 indicating that independent variables most likely influence the dependent variables.

5.3.2.2 Statistical tests of predictor variables (independent variables)

The statistical significance of the individual predictor (independent) variables will test using Wald's chi-square statistic, in which the predictor variable will be significant if its *p*-value is less than 0.05.

5.3.2.3 Hosmer and Leme show's test

This test will indicate the model's appropriateness.

5.3.2.4 Classification of observations

The model's predictive accuracy will present in a classification table, in which the predicted outcome (1/0) compared to the actual outcome (1/0). Hence, the classification of observations is critical in predicting the number of high (y = 1) and low (y = 0) claims. This classification considers a guide for underwriters, and specifically at times to renew individual health insurance policies.

5.4 Data Analysis and Discussion of Results

Walker (1998) and Edwards (2003) recommended a logistic model for data analysis due to its many appropriate characteristics, and specifically if the dependent variable is dichotomous and the independent variables are a mix of continuous and categorical variables. Further, Gebotys (2000) noted the logistic model as an important tool, as it provides an appropriate test for the coefficients' significance and the knowledge of the independent variable's effect on the dependent (dichotomous) variable. Moreover, the logistic model ranks the independent variables' effects. Hence, the researchers analyzed data using the SPSS version 21 statistical package, and considered the logistic model as appropriate for two reason. First, it does not require independent

calculate the regression coefficients.

The multivariate model indicated in Equation(1) is useful when the response variable is continuous, but is not appropriate for dichotomous response variables, as is the case when the claim amount Y is high (1)/low (0). Consequently, the multivariate model would not produce values restricted to one or zero, as we desire, as many uninterpretable values between zero and one and greater than one could be obtained. We recommend using a logistic regression model to prevent this problem, as this indirectly models the response variable based on probabilities associated with the values of Y.

Hence, we can use $\pi(x)$ to represent the probability that y = 1, which involves high claim amounts, and 1- $\pi(x)$ to be the probability that y=0,which involves low claim amounts(where the x in $\pi(x)$ is a vector representing the set of the independent predictor variables $X_1, X_2, X_3 \ldots X_n$).

We note these probabilities as

$$\pi(x) = P = P(Y=1 \mid X_1, X_2, X_3, \text{-------}, X_n) \quad (2)$$

$$1-\pi(x)=1- P= P(Y=0 \mid X_1, X_2, X_3, \text{-------}, X_n).$$

By taking the natural logarithm of the odds (log-odds) for Equation (2) to favor Y = 1, we obtain the following equation:

$$Ln\frac{\pi}{1-\pi} = ln\frac{P(Y=1|X_1,X_2,\ldots X_n)}{1-P(Y=1|X_1,X_2,\ldots X_n)} = \beta_0 + \sum_{j=1}^{n}\beta_j X_j \quad (3)$$

We use the inverse of the logit transformation of Equation (3) to arrive at the following equation, which represents our model:

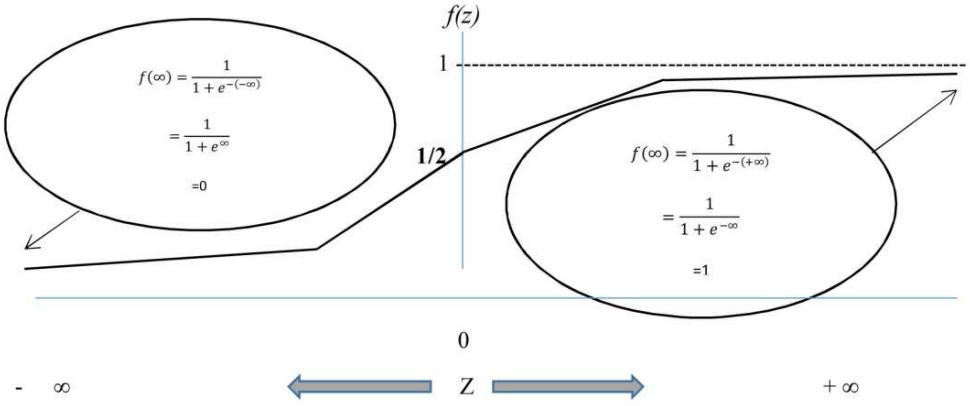
$$P(Y=1|X_1,X_2,\ldots X_n) = \frac{e^{\beta_0+\sum_{j=1}^{n}\beta_j X_j}}{1+e^{\beta_0+\sum_{j=1}^{n}\beta_j X_j}} = \frac{1}{e^{-(\beta_0+\sum_{j=1}^{n}\beta_j X_j)}} \quad (4)$$

We use Equation (4) to fit the logistic regression model to our data and calculate the conditional mean for claim amounts between 0 and 1, where

$$\text{Claim amount (a dichotomous variable)} = \begin{cases} 1 \; if \; the \; claim \; is \; high \\ 0 \qquad\qquad if \; other \end{cases}$$

Further, X_1, X_2, $X_3 \ldots$ X_n are the following predictor variables, respectively: the insured's age, gender, nationality, employment status, and marital status; P is the probability that the claim is high (Hosmer and Lemeshow2000; Rush2001; Yusuff, Ngah, and Yahaya, 2012; Achia, Wangombe, and Khadioli, 2010; Khodair, 2012; Strömbergsson, 2009; Al Farhood, 2014).

It is noteworthy that the logistic regression model indicated in Equation (4) is a continuous function ranging from zero to one, which approximates zero when the right side of the equation approximates -∞, and approximates 1 when the right side of the equation approximates +∞, as Figure (2) illustrates.



Range: $0 \leq f(z) \leq 1$

Source: Khodair, A. (2012).

[Fig. 2]

in the Saudi insurance market. The researchers collected the available data (different variables) from that large company for each insured, as the below Table (2) illustrates. The authors then used the SPSS software suite to calculate some descriptive statistics regarding the different data variables, as the following Table (2) indicates:

Table (2) Descriptive Statistics of Data

Variables	Mean	Std. Deviation	N
Claim Amount	2,482.26	9,945.426	1,658
Age	51.70	8.770	1,658
Gender	.31	.461	1,658
Nationality	.32	.466	1,658
Residence	.31	.464	1,658
Employment	.19	.396	1,658
Marital Status 1	.39	.488	1,658
Marital Status 2	.61	.488	1,658

By looking at Table (2),we find the average age of insureds 51for the following reasons:

a. The Council of the Cooperative Health Insurance obligated the insurance companies to accept all categories of the elderly and cancellation of the age requirement in the policy.
b. Most older insureds are more reluctant to hospital، so the random sample has a large proportion of retirees and the elderly insureds.
c. The cost of insurance increases when the age of the insured increased and therefore the underwriter must take sufficient care in the application of the principles of the underwriter to determine the appropriate cost.

Moreover, the researchers classified the Claim Amount variable into two categories: high claims, which are greater than the average claim (2,482.26 SR);and low claims, or less than average claims (2482.26 SR). Hence, the data consists of a claim amount, as a dependent dichotomous variable (high cost =1, low cost =0), and seven independent variables (Age, Gender, Nationality, Residence, Employment, Marital Status 1, and Marital Status2). As Table (3) demonstrates, all independent variables are binary variables except the age variable, which is continuous:

Table (3) Description of Variables

Variables	Definition	Characteristic
Age	**Insured's age**	Continuous
Gender	**Insured's gender**	Binary (1 if male, 0 if female)
Nationality	**Insured's nationality**	Binary (1 if Saudi, 0 if Non-Saudi)
Residence	**Insured's place of residence**	Binary (1 if inside Riyadh, 0 if outside Riyadh)
Employment	**Insured's occupation**	Binary (1 if employed, 0 if not employed)
Marital Status 1	**Insured's Marital Status**	(1single, married, others)

5.2 Model Propositions

This research attempts to investigate demographic factors' effects on individual health insurance claim amounts. It attempts to operationalize the claim not only in terms of the importance of each factor (means), but also in terms of the relative importance given to each factor; therefore, it is hypothesized that:

H1: Policyholders' demographic factors affect health insurance claim amounts.

H2: Demographic factors control the degree of risk for policyholders, as they differ from

One policyholder to the other.

5.3 The Model

To structure the logistic regression model and study the policyholders' demographic factors that affect individual health insurance claim amounts, we must first establish a fundamental model for a multiple regression analysis. The outcome variable in the multiple regression analysis is a linear combination of a set of predictors. We propose the following equation for outcome variable Y, and a set of *n* predictors variables X_1, X_2, X_3……..X_n:

$$Y= \beta_0+ \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 +\text{-------} + \beta_n X_n+ \varepsilon \quad (1)$$

Where β0 is the expected value of Y when the X is set to 0, β_j is the regression coefficient for each corresponding predictor variable X_j, and is the error in the prediction. Each regression coefficient β_j represents the slope of the regression line; the larger the β_j, the more influence the independent variable X has on the dependent variable Y and the difference between Y- = Y`, where Y` represents the expected value of Y (i.e., E(Y| X_1, X_2, X_3,-------, X_n). The ordinary least square (OLS) estimation may use to

logistic regression model had applied in many research fields. Consequently, the authors used retrospective data in China to identify the risk factors affecting medical service utilization. Overpek et al.'s (1997) logistic model noted that the dependent variable Y is the outcome regarding whether the insured uses a medical service. Specifically, Y= 1 indicates the insured is a hospital inpatient, and Y= 0 indicates no hospitalization. Further, X is the interpretation vector that demonstrates the impact of risk factors.

In recent study Mohammad et al (2018) has studied health service use among households with different socio-economic status in Isfahan province; and to investigate probable inequality determinants in service utilization. Lambon et al (2017) has investigated the factors that affect neonatal deaths as well as examine the effect of the Ghana Health Insurance on neonatal deaths in Ghana.

Hence, the previous studies concentrated on studying the risk factors and their effects on medical expenses and medical services, and not health insurance claims amounts. The Saudi insurance market has more insurance companies (about 34 companies), most of them are operating in health insurance, but the competition among them does not take the technical aspects of underwriting in consideration. Consequently, some of insurance companies incurred excessive losses in health insurance, and exited the market because these losses affected their financial position.

Moreover, The Saudi insurance institutions need to be highlighted for several reasons (https://www.mubasher.info/news/3072903)

1. At the end of 2016, 18 companies were unable to address their accumulated losses as they faced several difficulties in meeting the requirements of the solvency margins

2. Losses of insurance companies in both medical insurance and car insurance are very high and range from 70 to 75% of the market, and represent a major cause of losses

3. Insurance companies in the Saudi market still rely heavily on compulsory insurance products namely health insurance and vehicle insurance, which account for 84% of the size of the market.

4. There is a price war between companies and losses in financial results, which affected the insurance sector in particular medical insurance and car insurance

5. The trend of a number of insurance companies to merge recently due to weak opportunities for competition and erosion of capital.

Therefore, researchers believe it is necessary to study risk factors (i.e., the insured's demographic factors) and their effects on individual health insurance claims amounts in the Saudi insurance market. This study will help the underwriter to know the key determinants for health insurance claims in the kingdom of Saudi Arabia.

3. Research Objective

An analysis of risk factors and their effects on claims amounts and risk control in individual health insurance is important because it is necessary to identify these important risk factors that affect claims amounts. These factors will not only establish an underwriting strategy and renew insurance policies through moderate rates, but will also provide a standard technique for claim administration and health insurance management. Hence, this research will pursue the following objectives:

- To develop a statistical model for health insurance claims (i.e., high or low claims) in the Saudi health insurance market.
- To explore the important demographic factors affecting health insurance claims.
- To guide the underwriter in making optimal decisions before renewing individual health insurance policies in the Saudi insurance market.

Therefore, this study will attempt to answer the following questions:

- What are the significant demographic factors affecting health insurance claims and their predicted ability?
- What are the arrangement of demographic factors affecting health insurance claims according to their importance and their predicted ability(priority) toward health insurance claims (i.e., high or low claims)?

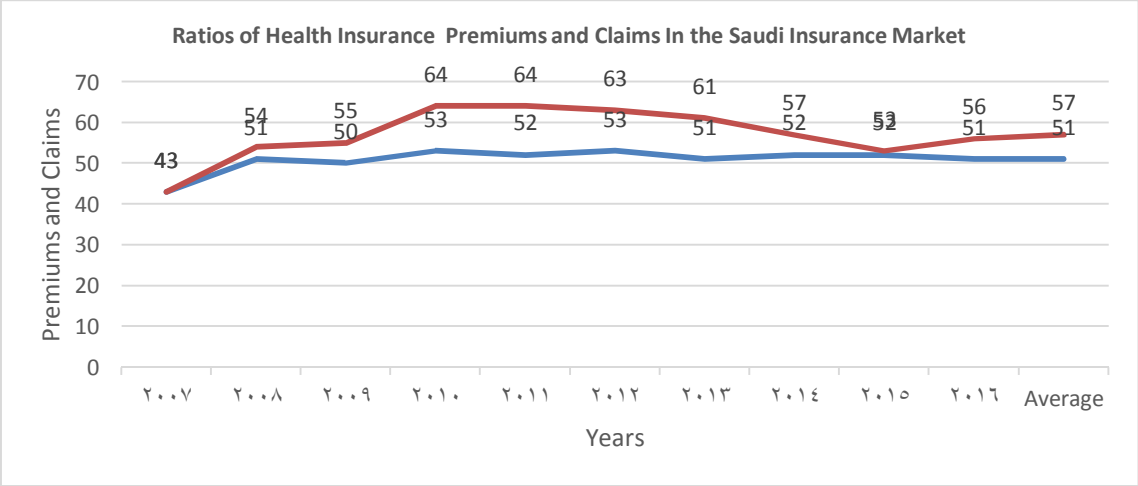
4. Importance of Research

This study endeavors to provide an effective tool for selecting potential insureds in the Saudi health insurance market, aside from risk control, and for renewing current policies at moderate rates.

5. Research Methodology

5.1 Data Description

By illuminating individual health insurance in the Saudi insurance market, we discover that many insurance companies provide this type of insurance. Therefore, the research population is all insureds with individual health insurance policies in the Saudi insurance market. The researchers have selected a simple random sample that consists of 1,658 insureds from a single insurance company, which considers the oldest and largest insurance companies



Source: Preparing by researchers using Excel.

Figure (1)

We consider both Table (1) and Figure (1) to discover that the ratios of health insurance claims are always higher than ratios of premiums in the Saudi insurance market. Consequently, some health insurance companies in the Saudi insurance market incurred heavy losses and exited the market, as these losses impacted their financial position (Ali I.A see: http://www.al-sharq.com/news). Therefore, insurance companies must apply the proper underwriting procedures to manage the risks in adverse selection. The identification and measurement of risk factors affecting medical costs has become critical in establishing an underwriting strategy. Although health insurance in Saudi Arabia is compulsory, we must discover the factors that affect the utilization of medical services. The knowledge of these factors are very important forth distribution of medical costs among the insured, which will balance funds and encourage insurance companies and medical service providers to continue in the insurance market. Moreover, an analysis of the risk factors that affect health insurance claims is useful in controlling risk, managing the risk, in particular for, adverse selection. As well, to compel the insurance companies to apply proper underwriting procedures. A knowledge of health insurance claim determinants will help the underwriter to accept risk coverage, with normal rates if the claim is normal (not high), or accept with an increased rate if the claim is abnormal (high).

Hence, this paper will attempt to explore the determinants of health insurance, which affect claims in the Kingdom of Saudi Arabia.

2. Literature Review

Seng et al. (1992), Chen and Zhao (1994), and Zhang (1996) have analyzed the risk factors of medical service expense. These authors divided the insured that utilize medical services into different groups, with different levels of suspicious risk factors, and then compared each group's average medical service expenses to identify the risk factors. Chen (2003) analyzed the risk factors of medical expenses in China using multivariable statistical models, such as multi-linear regression models. Ismail (2016) used such health risk factors as age, sex, nationality, marital status, occupation, and place of residence, to classify the health risks in the Saudi insurance market: high, normal, moderate, and bad risks. Further, Ismail et al. (2013b) developed a quantitative model to finance the net medical insurance costs for low-income people in the Kingdom of Saudi Arabia. Their model depended on an aggregate probability distribution, and suggested a distribution for the costs of low-income medical care between the Saudi government and the low-income citizens. As well, Ismail et al. (2013a) have been studied the challenges of medical insurance in Saudi Arabia, and identified the most important challenges that Saudi Arabia's medical insurance system faces in its development. They ranked these challenges according to their average importance, and recommended that some modifications should be made to the cooperative medical insurance policy.

Moreover, Scott (2010) studied U.S. health care reform and the potential effects of health insurance. Overpek et al. (1997) further mentioned that the

A Quantitative Model to Identify Key Determinants for Health Insurance Claims in the Kingdom of Saudi Arabia

Mohamed Abdelmawla Osman
Finance Dep., King Saud University
Insurance & Statistics Dep., Tanta University
maosman@ksu.edu.sa

Emad Abdelgalil Ali Ismail
Quantitative Analysis Dep., King Saud University
Insurance& Mathematics Dep., Cairo University
emadali@ksu.edu.sa

Keyword: health insurance, claim amount, logistic model, demographic factors, Saudi Arabia.
Abstract: The insurance sector in the Kingdom of Saudi Arabia has been growing in recent year in particular, the fields of health and automobile insurance. However, some insurance companies have incurred excessive losses in health insurance and exited the market as their financial position affected. Therefore, this paper explores the determinants of individuals' health insurance that affect their claims. In other words, what are the demographic factors for policyholders that affect their health insurance claim amounts in the Kingdom of Saudi Arabia? An analysis of the risk factors that affect such claims is useful in risk control, as well as in managing the risks of adverse selection and assisting insurance companies to apply proper underwriting procedures. Some statistical models may use to analyze the determinants of health insurance claims; the paper suggests a statistical (logistic) model for the primary determinants of these claims. This model revealed five significant demographic factors: age, gender, nationality, and two types of marital status factors may affect the classification of individual health insurance claims by an overall percentage 90.8%.These demographic factors may use as a guide for underwriters in making crucial and successful decisions while rating new health insurance policies or renewing current policies.

1. Introduction

Individual medical insurance is important in providing health security to individuals and families, as this insurance covers the medical expenses incurred during illness. Simultaneously, the insured must pay sufficient premium to the insurer to obtain indemnity from medical expenses when injury or illness occurs. The conflict that exists among the insurer, the insured, and the medical service provider may cause ***medical expenses*** beyond expectations. Medical expenditures and claim amounts influence not only by the insured's age or gender, but also by other factors. Zhang (1996) notes that it is necessary to identify the risk factors in medical expenses and quantitatively describe their influence in order to control health insurance claim amounts. Many risk factors can affect medical expenses, and they divide into three categories: the factors that affect medical service utilization, those affecting medical service costs, and those affecting both (Chen 2003). The first factors can increase the probability of utilization, but the second can increase the amount of service. Hence, identifying these factors is critical in health insurance risk control.

Individual medical insurance is compulsory in the Kingdom of Saudi Arabia. Table 1 illustrates the premiums of this type of insurance represents 51% of the Saudi insurance market at the end of 2016,but the claims represents 56% (Saudi Arabian Monetary Agency, 2016).

Table (1) Ratios of Health Insurance Premiums and Claims in the Saudi Insurance Market, 2007 to 2016 (%)

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Average Ratio*
Premiums (1)	43	51	50	53	52	53	51	52	52	51	51
Claims (2)	43	54	55	64	64	63	61	57	53	56	57
Percentage (2) ÷(1)	100	106	110	121	123	119	120	110	102	110	112

Source: Saudi Arabian Monetary Agency (SAMA).Over period 2007- 2016.