UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

BIOSTATISTICS II

ASSIGNMENT 2

2022

STUDENT NAME: Nokwanda Themba

STUDENT NO.: THMNOK003

COURSE CODE: PPH7092S

DUE DATE: 12 September 2022

THIS ASSIGNMENT INCLUDES::

1. **Software used: R**
2. **Data Set: BiostatII_2022_Assign2_data.51.csv**
3. **THMNOK003_R Script Assignment 2**

LIST OF APPENDICES:

I. **R Script Code File**

Total [55 Marks]

Question 1 [6 marks]: **Odd ratio calculations (by hand or using the software)**

Calculate and present crude odds ratios for the relationship: perceived state of health (perceived_health) vs medical aid scheme change (med_scheme).

Provide a labelled frequency table (Table 1) including the margins and interpreting the OR in a short paragraph. [6]

Table 1: Contingency table for the relationship between perceived state of health and medical aid scheme change

Exposure (perceived state of health)		Outcome (medical aid scheme change)		
		change	**no change**	**Total:**
	good	96	325	421
	poor	18	91	109
	Total:	114	416	530

Odds ratio calculations:

- Odds of medical aid scheme change among those who had good perceived state of health:

$= 96 / 325 = 0.295 = 0.30$

The odds for medical aid scheme change in individuals who had good perceived state of health is 0.30

- Odds of exposure medical aid scheme change among those who had poor perceived state of health:

$= 18 / 91 = 0.198 = 0.20$

The odds for medical aid scheme change in individuals who had poor perceived state of health is =0.20

Crude Odds ratio:

$= 0.30/0.20 = 1.49$ **OR** $= (a\,x\,d) / (c\,x\,b) = (96\,x\,91) / (18\,x\,325) = 1.49$

- **Measures of association strength:**

The crude odds ratio **(OR)** is **1.49 (95% CI 0.86 *to* 2.6)**. We can state that the odds for medical aid scheme change are higher in individuals who had a good perceived state of health, since OR > 1.

Interpretation of crude OR:

Individuals who had a good perceived state of health have a 1.49 higher odds of changing their medical aid scheme, compared to individuals who had a poor perceived state of health.

Question 2 [14 marks]: Univariable logistic regression

a) Present a single table (Table 2) showing the results of the models below with estimated coefficients (on both log-odds and odds ratio scale) and 95% confidence intervals on odds ratio scale.

i) Univariable logistic models with the covariates perceived state of health (perceived_health) and charges. [8]

Table 2: Univariate logistic regression analysis on the association of perceived state of health and charges variables with medical aid scheme changes

	Univariate analysis					
	Model 1			Model 2		
Variables	β	**OR (95% CI)**	**p-value**	β	**OR (95% CI)**	**p-value**
perceived state of health (poor)	-0.40	0.67 (0.37 - 1.14)	0.2		-	-
charges (rands)	-	-	-	0.00	1.00 (1.00-1.00)	<0.001

OR: Odds ratio, CI: Confidence interval, β: Beta coefficient

b) In no more than 300 words, describe your findings from the models you fitted in 2a(i) and include interpretations of the associations found. [6]

Model 1: Perceived health on Medical aid scheme change

Univariate analysis with a categorical predictor (perceived state of health):

- **Conceptual equation:**

$log\ (p/1\ -\ p)\ =\ \beta0\ +\ \beta1X1$

$log\ (p/1\ -\ p)\ =\ \beta0\ +\ \beta1\ x\ perceived\ state\ of\ health$

p = probability of medical aid scheme change (Y=1; if there was a change)

Note: We code $X1$=1 if an individual's perceived state of health is 'poor';then log ($p1$/1-$p1$) = $\beta0$ + $\beta1$

We code $X1$=0 if an individual's perceived state of health is 'good"; then log ($p0$/1-$p0$)= $\beta0$

- **Fitted regression equation:**

$log\ (\hat{p}\ \ (medical\ aid\ scheme\ change)/1\ -\ \hat{p}\ (medical\ aid\ scheme\ changes)\ =\ \beta0\ +\ \beta1\ x\ perceived\ state\ of\ health$

$=\ \ -\ 1.22\ +\ (-\ 0.401)\ x\ perceived\ state\ of\ health$

- **Model fit statistics:** From Table 2: the odd ratio is **0.67** with **95% CI** being **0.37 and 1.14. p-value =0.2**

Interpretation: There appears to be no statistical significant association between perceived health and medical aid scheme changes (p > 0.05).Perceived health is not a statistically significant predictor to medical aid scheme change p = 0.2 (p > 0.05).

Calculating Beta's from OR's β1= log$e$ (OR) or ln (OR)

β1= log$e$ (0.67) or ln(0.67) = -0.40 (the OR associated with poor perceived state of health is exp (-0.40) = 0.67)

An individual with a poor perceived state of health has 0.67 (0.37, 1.14) decreased odds (OR < 1)of changing their medical aid scheme compared to an individual with a good perceived state of health.
The odds of changing your medical aid scheme for individuals with a poor perceived state of health is ((1-0.67)x100) 33% less likely as compared to individuals with a good perceived state of health.

On average, the 'odds' of medical aid scheme change for those who had a poor perceived state of health are 0.67 times the 'odds' of medical aid scheme change for those who had a good perceived state of health. This association is however, not statistically significant (p-value >.05)

For Individuals who had a good perceived state of health: $1/OR\ =\ 1/0.67\ =\ 1.49$

Interpretation: Individuals who had a good perceived state of health have increased 'odds' of medical aid scheme change (OR > 1) compared to those who had a poor perceived state of health. The odds of medical aid scheme change for individuals who had a good perceived state of health are almost 2 times more likely than those for individuals who had a poor perceived state of health.

Interpretation (log odds)

- ➢ **Perceived state of health (poor):** ˆβ1 = -0.40

The estimated log odds of medical aid scheme change are 0.40 times lower in individuals who have a poor perceived state of health (versus those who had a good perceived state of health), keeping other covariates constant.i.e having the same age of primary beneficiary and in the same spouse group. This effect is not significant (p-value>.05)

Model 2: Charges on Medical aid scheme change

Univariate analysis with a continuous predictor (charges):

- **Conceptual equation:**

$log\ (p/1\ -\ p)\ =\ \beta0\ +\ \beta1X1$

$log\ (p/1\ -\ p)\ =\ \beta0\ +\ \beta1\ x\ charges$

p = probability of medical aid scheme change (Y=1; if there was a change)

- **Fitted regression equation:**

$log\ (\hat{p}\ \ (medical\ aid\ scheme\ change)/1\ -\ \hat{p}\ (medical\ aid\ scheme\ change))\ =\ \beta0\ +\ \beta1\ x\ charges$

$=\ 5.83\ +\ 0.00\times charges$

- **Model fit statistic:** From Table 2, the **odd ratio is 1.00**, with **95% CI being 1.00 and 1.00. p-value<0.001**

Interpretation: There appears to be a statistical significant association between charges and medical aid scheme changes (p < 0.05). Charges appears to be a statistically significant predictor to medical aid scheme change p <0.001 (p <0.05)

- **Calculating Beta's from OR's $\beta 1$= log$_e$ (OR) or ln (OR)**

$\beta 1$= log$_e$ (1. 00) or ln(1. 00) = 0.00 (the OR associated with charges is exp (0.00) = 1. 00)

On average, a unit increase in charges has no difference in the 'odds' of medical aid scheme change (OR = 1).For every 1 rand increase in charges, the odds of an individual changing medical aid schemes is approximately 0. Therefore, there is no association between the medical aid scheme change and charges. This association is statistically significant p<0.001 (p-value <.05)

Interpretation (log odds)

➢ **Charges:** ˆβ2 =0.00

A unit increase in charges is associated with no increase in the estimated log odds of medical aid scheme change, keeping other covariates constant. This effect is statistically significant (p- value<.05)

Question 3 [15 marks]: Multivariable logistic regression

Develop a multivariable model including spouse and age. Explore whether age and spouse modify the effects of perceived state of health (perceived_health) or charges. Present your final model(s) in a table (Table 3), with odds ratios and their 95% CIs [5].

Next, explain the process, rationale and findings from your analysis. In particular, explain the association of age and spouse on medical aid scheme changes. [10]

Table 3: Multivariate logistic regression analysis on the association of variables age and spouse on perceived state of health (model 1) and charges (model 2) variables with medical aid scheme changes

	Multivariate analysis					
	Model 1			**Model 2**		
Variables	β	**OR (95% CI)**	**p-value**	β	**OR (95% CI)**	**p-value**
spouse (yes)	1.08	2.94 (0.93-9.62)	0.069	1.85	6.38 (2.70-16.4)	<0.001
age (years)	1.82	6.15 (4.27-10.03)	<0.001	1.83	6.21 (2.55-17.2)	<0.001
perceived state of health (poor)	- 1.38	0.25 (0.06-0.95)	0.047	-	-	-
charges (rands)		-	-	0.00	1.00 (1.00-1.00)	0.91

OR: Odds ratio, CI: Confidence interval

Multivariate Model 1: spouse and age effect on perceived health

Multivariate analysis with a categorical predictor (perceived state of health):

- **Model fit statistics:** From Table 3, the **OR** is **0.25** with **95% CI** being **0.06 and 0.95**, **β=-1.38, p-value =0.047**

Interpretation: Perceived health is a statistically significant predictor of medical aid scheme change p = 0.047 (p < 0.05, keeping all other covariates constant. There appears to be a statistical significant association between perceived health and medical aid scheme changes (p < 0.05). We are also 95% confident that the true odds ratio is between 0.06 and 0.95

The odds of changing your medical aid scheme for individuals with a poor perceived state of health is ((1-0.25) x100) 75% less likely as compared to individuals with a good perceived state of health.

Individuals with a poor perceived state of health have 0.25 (0.06, 0.95) decreased odds (OR < 1) of medical aid scheme change compared to individuals with a good perceived state of health, adjusted for spouse and age. This association is statistically significant (p- value<.05)

On average, the 'odds' of medical aid scheme change for those who had a poor perceived state of health are 0.25 times the 'odds' of medical aid scheme change for those who had a good perceived state of health. This association is statistically significant (p-value<.05)

Multivariate Model 2: spouse and age effect on charges

Multivariate analysis with a continuous predictor(charges):

- **Model fit statistic:** From Table 3, the **OR is 1.00**, with **95% CI being 1.00 and 1.00**, **β=0, p-value=0.91**

Interpretation: There appears to be no statistical significant association between charges and medical aid scheme changes ($p > 0.05$). Charges are not a statistically significant predictor to medical aid scheme change, keeping all other covariates constant. p=0.91 ($p > 0.05$)

On average, a unit increase in charges has no difference in the 'odds' of medical aid scheme change (OR = 1).For every 1 rand increase in charges, the odds of an individual changing medical aid schemes is approximately 0. Therefore, there is no association between the medical aid scheme change and charges. This association is not statistically significant (p-value >.05)

Model choice: multivariate model 1-perceived health

The charges predictor variable appears to be not statistically significant p=0.9 ($p>0.05$) in our model (multivariate model 2), the OR is 1.00 and the 95% Confidence Interval is between 1.00-1.00, which is entirely too narrow. Although a narrow 95% CI is good for precision, a confidence interval that is too narrow implies that there is a smaller chance of obtaining more precise population estimates within that interval. We would not be able to find any true odds ratio values between 1.00 and 1.00. An OR=1 which indicates that there is no difference in unit change and no association between the medical aid scheme change and charges. This model would not give us accurate predicted (estimated) probability values, and would not help us in predicting the outcome.

Perceived state of health would be a better predictor variable in our model (multivariate model 1). It is statistically significant p = 0.047 ($p < 0.05$), the OR is 0.25 (OR < 1) which tells us that there is a reduced probability of outcome medical aid scheme changes in individuals with poor perceived state of health compared to individuals with good perceived state of health, the 95% confidence interval is well in a good narrow range of between 0.06 and 0.95. This means that our true odds ratio could lie between those given values, giving us a greater degree of precision without it being too narrow (as opposed to the 95% CI of the charges predictor variable in multivariate model 2) thus enabling us to make more precise population estimates.

Therefore, multivariate Model 1 is a better choice moving forward. The odds ratio, effect size and power is much higher than multivariate model 2, and the confidence interval is narrow enough to give us accurate results. It would help us in predicting probability estimates of the outcome.

For Multivariate Model 1: spouse and age effect on perceived health

- **Conceptual equation:**

$$log\,(p/1 - p) = \beta 0 + \beta 1\, X1 + \beta 2\, X2 + \beta 3\, X3$$

p = probability of medical aid scheme change (Y=1; if there was a change); holding the other predictor variables constant at a certain value

- **Fitted regression equation:**

$$log\,(\hat{p}\ (medical\ aid\ scheme\ change)/1 - \hat{p}\,(medical\ aid\ scheme\ changes)) = \hat{}\beta 0 + \hat{}\beta 1\ x\ spouse + \hat{}\beta 2\ x\ age + \hat{}\beta 3\ x\ perceived\ state\ of\ health$$

$$log\,(\hat{p}\ (medical\ aid\ scheme\ change)/1 - \hat{p}\,(medical\ aid\ scheme\ changes)) = -\ 62.4 + 1.08\ xspouse + 1.82\ x\ age - 1.38\ x\ perceived\ state\ of\ health$$

Using Wald test to test whether predictor variables "spouse" and "age" are useful in our model:

$H0$: βk = 0 Some set of predictor variables are all equal to zero

$H0$: $\beta k \neq 0$ Not all predictor variables in the set are equal to zero

- **Model fit statistics**: p<0.001

The chi-squared test statistic of 72.3, with two degrees of freedom indicates better predictive power for a set of variables. This is associated with a p-value of <0.001 indicating that the overall effect of rank is statistically significant($p < 0.05$).

The estimate (β) significantly improves model fit and the variable is significant. Since this p-value is less than .05 ($p < 0.05$), we have enough statistical evidence to reject the null hypothesis of the Wald test. Not all predictor variables in the set are equal to zero. This means we can assume the regression coefficients for the predictor variables "spouse" and "age" are not equal to zero. Therefore, we should include all the variables in our model since they statistically significantly improve the overall fit of the model.($p < 0.05$) and do affect the model in any meaningful way.

- For spouse we have $H0$: $\beta 1 \neq 0$ (i.e., spouse is related to the outcome (i.e. medical aid scheme changes)
- For age we have $H0$: $\beta 2 \neq 0$ (i.e., age is related to the outcome (i.e. medical aid scheme changes)
- For perceived health we have $H0$: $\beta 3 \neq 0$ (i.e., perceived state of health is related to the outcome (i.e. medical aid scheme changes)

Additionally, from the model output, we have reported: Null deviance: 551.87 on 529 degrees of freedom; Residual deviance: 154.77 on 526 degrees of freedom

The residual deviance is much lower than the null deviance, which means that our predictor variables are useful/significant in the model.

Interpretation (log odds)

- **Perceived state of health (poor):** $\hat{}\beta 3$ = -1.38

The estimated log odds of medical aid scheme change are 1.38 times lower in individuals who have a poor perceived state of health (versus those who had a good perceived state of health), keeping other covariates constant.i.e having the same age of primary beneficiary and in the same spouse group. This effect is not significant (p-value>.05)

➢ **Spouse (yes):** $\hat{\beta}1 = 1.08$

The estimated log odds of medical aid scheme change are 1.08 times higher in individuals who have a spouse (versus those who did not have a spouse), keeping other covariates constant i.e. having the same age of primary beneficiary and in the perceived state of health group. This effect is not significant (p-value>.05)

➢ **Age of primary beneficiary:** $\hat{\beta}2 = 1.82$

A unit increase in age is associated with an increase of 1.82 in the estimated log odds of medical aid scheme change, keeping other covariates constant. This effect is statistically significant (p- value<.05)

Interpretation (odds):

➢ **Poor perceived state of health**

The OR associated with poor perceived state of health is exp(-1.38) = 0.25. Individuals who had a poor perceived state of health have decreased 'odds' of medical aid scheme change (OR < 1) compared to those who had a good perceived state of health . On average, the 'odds' of medical aid scheme change for those who have a poor perceived state of health are 0.25 times the 'odds' of medical aid scheme change for those who had a good perceived state of health (100×(1-0.25) = 75% decrease in odds), keeping other covariates constant. This association is however not statistically significant (p-value >.05)

➢ **Spouse (yes):**

The OR associated with having a spouse is exp(1.08) = 2.94. Individuals who have a spouse have increased 'odds' of medical aid scheme change (OR > 1) compared to those who did not have a spouse. On average, the 'odds' of medical aid scheme change for those who have a spouse are 2.94 times the 'odds' of medical aid scheme change for those who do not have a spouse, keeping other covariates constant. This association is however not statistically significant (p-value >.05)

➢ **Age of primary beneficiary:**

The OR associated with age at admission is exp(1.82) = 6.17. On average, a unit increase in age increases the 'odds' of medical aid scheme change 6 times, keeping other covariates constant. This association is statistically significant (p-value <.05)

Question 4 [7 marks]: Predictions (Using final fitted model in Q3)

- **Conceptual equation:**

$log\ (p/1 - p) = \beta 0 + \beta 1\, X1 + \beta 2\, X2 + \beta 3\, X3$

p = probability of medical aid scheme change (Y=1; if there was a change); *holding the other predictor variables constant at a certain value*

- **Fitted regression equation:**

$log\ (\hat{p}\ (medical\ aid\ scheme\ change)/1 - \hat{p}\ (medical\ aid\ scheme\ changes)) = \hat{\beta}0 + \hat{\beta}1\ x\ spouse + \hat{\beta}2\ x\ age + \hat{\beta}3\ x\ perceived\ state\ of\ health$

$log\ (\hat{p}\ (medical\ aid\ scheme\ change)/1 - \hat{p}\ (medical\ aid\ scheme\ changes)) =- 62.4 + 1.08\ xspouse + 1.82\ x\ age$

$- 1.38\ x\ perceived\ state\ of\ health$

a) Compute the predicted (estimated) probability of changing medical aid schemes (specify the values of the covariate combination used to obtain the probability). [2]

Prognostic/covariate combination values:

- Median Age: 32 years of age
- Spouse-Yes
- Perceived state of health-poor

Outcome of interest: medical aid scheme change

Note: Perceived state of health is coded as poor=1 (reference); good=0 *i.e. b0 + b1 if individual has poor perceived state of health; bo if individual is good perceived state of health.* Spouse is coded as yes=1(reference); no=0 *i.e. b0 + b1 if individual has a spouse; bo if individual does not have a spouse*

"What is the estimated probability of a medical aid scheme change for a median age individual with a spouse who has a poor perceived state of health?"

- ***Fitted equation:***

$log\ (\hat{p}\ (medical\ aid\ scheme\ change)/1 - \hat{p}\ (medical\ aid\ scheme\ changes)) = \hat{\beta}0 + \hat{\beta}1\ x\ spouse + \hat{\beta}2\ x\ age + \hat{\beta}3\ x\ perceived\ state\ of\ health$

$log\ (\hat{p}\ (medical\ aid\ scheme\ change)/1 - \hat{p}\ (medical\ aid\ scheme\ changes)) =- 62.4 + 1.08\ x\ spouse + 1.82\ x\ age$

$- 1.38\ x\ perceived\ state\ of\ health$

Solving for $\hat{p}$ $(medical\ aid\ scheme\ change)$:

$\hat{p}$ $(medical\ aid\ scheme\ change)$

$= exp\,(-\ 62.4\ +\ 1.08\ x\ spouse_{yes}\ +\ 1.82\ x\ age\ -\ 1.38\ x\ perceived\ state\ of\ health_{poor})\ /$

$1\ +\ exp(-\ 62.4\ +\ 1.08\ x\ spouse_{yes}\ +\ 1.82\ x\ age\ -\ 1.38\ x\ perceived\ state\ of\ health_{poor})$

- **Dummy variable substitution:**

$\hat{p}$ $(medical\ aid\ scheme\ change)$

$= exp\,(-\ 62.4\ +\ 1.08\ x\ 1\ +\ 1.82\ x\ 32\ -\ 1.38\ x\ 1)\ /$

$1\ +\ exp(-\ 62.4\ +\ 1.08\ x\ 1\ +\ 1.82\ x\ 32\ -\ 1.38\ x\ 1)\ =\ 0.02312473\ =\ 0.02$

The estimated probability of a medical aid scheme change for a median age individual with a poor perceived state of health who has a spouse is 0.02

b) Use a classification that an individual will change their medical aid scheme if the predicted probability of change is 0.5 or greater. Tabulate and present the predicted versus actual medical aid changing in a labelled 2x2 table (Table 4).

Table 4: Confusion matrix for the accuracy prediction of our model:

		Actual values	
		Positive (1)	**Negative (0)**
Predicted values	**Positive (1)**	**100**	**24**
	Negative (0)	**14**	**392**

Calculate and present the proportion of the cases that are predicted correctly. [5]

Sensitivity = TP / TP + FN = 100/100+14 = **0.88; Specificity = TN / TN + FP** = 392 / 392 + 24 = 0.94

Accuracy = (TP+TN)/(TP+FP+TN+FN) = $(100\ +\ 392)/(100\ +\ 24\ +\ 392\ +\ 14)\ =\ 0.92$

The accuracy score reads as 92% for the given data and observations. From all the cases (positive and negative), 0.92 were predicted correctly.

Question 5 [10 marks]: Discussion. Summarise your final conclusions in less than 400 words. Justify your claims by backing them up with your results. [10 marks]

We looked at different factors that increase the likelihood of an individual changing their medical aid scheme, of which the main exposure was the individual's perceived state (good or poor). Other potential factors were an individual's monthly cost billed by medical aid scheme (charges), whether an individual has a spouse or not, the sex of primary beneficiary (male or female) and the age of primary beneficiary.

From univariate model 1, we can see that perceived state of health alone has an effect on the outcome (OR=0.67;95% CI (0.37- 1.14)). However in this model, perceived health is not a statistically significant predictor to medical aid scheme change p = 0.2 (p > 0.05). From univariate model 2, we can see that charges have less of an effect on an individual changing their medical aid scheme (OR=1.00; 95% CI (1.00-1.00)) indicating no difference in unit change. Although the model is statistically significant p<0.001; p<.05.

When we adjust for age and spouse, we see a significant effect on perceived health and charges with the outcome. We concluded that it would be important to include these factors in our model, since they have a significant effect ($\beta k \neq 0$), improving the overall fit of the model. Since including statistically significant predictors should lead to better predictions (i.e., better model fit) we can conclude that including age and spouse results in a statistically significant improvement in the fit of the model.

In our multivariate model 1, after adjusting for age and spouse, we can see that perceived state of health still has an effect on the outcome (OR=0.25 ; 95% CI (0.06, 0.95)) and is a statistically significant predictor of medical aid scheme change p = 0.047 (p < 0.05). Similarly, in our multivariate model 2, charges has no effect on our outcome (OR= 1.00; 95% CI (1.00-1.00)) and appears to have no statistically significant association with medical aid scheme changes, keeping all other covariates constant p=0.91 (p >0.05). With a significant odds ratio, large effect size,enough power and narrow confidence interval, perceived state of health is an overall better predictor variable in our multivariate model, thus enabling us to make more precise predicted probability estimates of the outcome.

In conclusion, from this model we were able to predict probabilities, as well as obtain predicted versus actual medical aid changing values in order to assess how accurate our model is in predicting the outcome. Our model indicated 92% of cases were predicted correctly. We can therefore conclude that perceived state of health is a sufficient indicator of whether an individual will change their medical aid scheme, keeping other factors such as age and spouse constant and in mind.