Solving real-world problems with statistical modelling

Cynthia Thinwa

Problem 1

As part of a larger case-control study, an investigator decided to identify factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams).

Women were matched according to age so that each case was matched to one control.

The dataset is **lowbwtmat.csv** and it contains the following variables:

Variable	Code/Value	Abbreviation
Identification Code		ID
Low Birth Weight	0= if weight is >=2500g 1=if weight is < 2500g	LOW
Mother's weight at last menstrual		LWT
Smoking during pregnancy	0=No 1=Yes	SMOKE
History of premature labour	0=No 1=Yes	PTL
History of hypertension	0=No 1=Yes	HT
Urinary tract infection	0=No 1=Yes	UI
Physician Visit during the first trimester	0=No 1=Yes	FTV

(i) For each variable given except for the outcome, state whether it is:

- a predictor,
- a confounder,
- an effect modifier or
- none of the above.

Give reason for your answer.

(ii) Fit the appropriate model based on part (i) and give a brief report of the analysis.

Solution for Question 1

(i)

Variable	Correct Category	Reason
Identification Code	None	The variable is simply for identification purposes.
Low Birth Weight	None	This is the variable that is of interest (the dependent variable), and is an outcome in and of itself.
Mother's weight at last menstrual	Confounder	This could be tied to the baby's birth weight in some way, but it is not directly related with the dependent variable.
Smoking during pregnancy	Predictor	This variable has a significant effect on the baby during pregnancy, therefore it is linked to the baby's birth weight.
History of premature labour	Confounder	Premature birth often results in a baby's birth weight being low, therefore this history could indicate that a premature birth could occur.
History of hypertension	Predictor	A history of hypertension could indicate that the baby's development was affected by hypertension, therefore it is linked to the baby's birth weight.
Urinary tract infection	Confounder	This could affect delivery of the baby, but it is not directly related to the dependent variable.
Physician Visit during the first trimester	Effect modifier	This variable has a significant effect on the baby during pregnancy, therefore it is linked to the baby's birth weight.

(ii)

An appropriate model to fit the data is the binary logistic regression model because the dependent variable has two categories. Stepwise regression based on the AIC was conducted, resulting in the following models:

```
## Start:  AIC=157.26
## low.birth.weight ~ 1
##
##                                Df Deviance    AIC
## + smoked                        1   147.94 151.94
## + history.premature.labour      1   148.86 152.86
## + UTI                           1   151.28 155.28
## + mothers.weight.last.menstrual 1   152.25 156.25
## <none>                              155.26 157.26
## + history.hypertension          1   153.46 157.46
## + doc.visit.first.trimester     1   154.36 158.36
##
## Step:  AIC=151.94
## low.birth.weight ~ smoked
##
```

```
##                                 Df Deviance    AIC
## + history.premature.labour       1   143.18 149.18
## + UTI                            1   144.10 150.10
## + mothers.weight.last.menstrual  1   144.70 150.70
## <none>                               147.94 151.94
## + history.hypertension           1   145.97 151.97
## + doc.visit.first.trimester      1   146.97 152.97
## - smoked                         1   155.26 157.26
##
## Step:  AIC=149.18
## low.birth.weight ~ smoked + history.premature.labour
##
##                                 Df Deviance    AIC
## + UTI                            1   140.11 148.11
## + mothers.weight.last.menstrual  1   140.34 148.34
## <none>                               143.18 149.18
## + history.hypertension           1   141.19 149.19
## + doc.visit.first.trimester      1   141.43 149.43
## - history.premature.labour       1   147.94 151.94
## - smoked                         1   148.86 152.86
##
## Step:  AIC=148.11
## low.birth.weight ~ smoked + history.premature.labour + UTI
##
##                                 Df Deviance    AIC
## + history.hypertension           1   137.28 147.28
## + mothers.weight.last.menstrual  1   137.74 147.74
## <none>                               140.11 148.11
## + doc.visit.first.trimester      1   138.40 148.40
## - UTI                            1   143.18 149.18
## - history.premature.labour       1   144.10 150.10
## - smoked                         1   145.94 151.94
##
## Step:  AIC=147.28
## low.birth.weight ~ smoked + history.premature.labour + UTI +
##     history.hypertension
##
##                                 Df Deviance    AIC
## + mothers.weight.last.menstrual  1   132.40 144.40
## <none>                               137.28 147.28
## + doc.visit.first.trimester      1   135.75 147.75
## - history.hypertension           1   140.11 148.11
## - history.premature.labour       1   141.19 149.19
## - UTI                            1   141.19 149.19
## - smoked                         1   143.13 151.13
##
## Step:  AIC=144.4
## low.birth.weight ~ smoked + history.premature.labour + UTI +
##     history.hypertension + mothers.weight.last.menstrual
##
##                                 Df Deviance    AIC
## <none>                               132.40 144.40
## + doc.visit.first.trimester      1   131.69 145.69
## - history.premature.labour       1   135.81 145.81
```

```
## - UTI                            1   135.95 145.95
## - mothers.weight.last.menstrual  1   137.28 147.28
## - history.hypertension           1   137.74 147.74
## - smoked                         1   138.41 148.41

##
## Call:  glm(formula = low.birth.weight ~ smoked + history.premature.labour +
##     UTI + history.hypertension + mothers.weight.last.menstrual,
##     family = binomial)
##
## Coefficients:
##                   (Intercept)                         smoked1
##                        1.0816                          1.0473
##     history.premature.labour1                            UTI1
##                        0.9635                          1.0747
##         history.hypertension1  mothers.weight.last.menstrual
##                        1.8132                         -0.0161
##
## Degrees of Freedom: 111 Total (i.e. Null);  106 Residual
## Null Deviance:       155.3
## Residual Deviance: 132.4     AIC: 144.4
```

A more detailed description of the best model findings were as follows:

```
##
## Call:
## glm(formula = low.birth.weight ~ smoked + history.premature.labour +
##     UTI + history.hypertension + mothers.weight.last.menstrual,
##     family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.95660  -0.88412  -0.07444   0.99778   1.75610
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    1.081571   0.979176   1.105   0.2693
## smoked1                        1.047292   0.434280   2.412   0.0159 *
## history.premature.labour1      0.963465   0.534113   1.804   0.0713 .
## UTI1                           1.074704   0.586632   1.832   0.0670 .
## history.hypertension1          1.813185   0.848591   2.137   0.0326 *
## mothers.weight.last.menstrual -0.016100   0.007609  -2.116   0.0344 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155.26  on 111  degrees of freedom
## Residual deviance: 132.40  on 106  degrees of freedom
## AIC: 144.4
##
## Number of Fisher Scoring iterations: 4
```

Statistical significance of the model

A Likelihood Ratio Test was conducted and the results were as follows:

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
111	155.2650	NA	NA	NA
106	132.4005	5	22.86446	0.0003583

Based on the results, the model is statistically significant because the p-value associated with the test statistic is « 0.05

Statistical significance of the coefficients

The mother's weight at her last menstrual, smoking during pregnancy and the mother having a history of hypertension were found to be statistically significant because their p-values were < 0.05.

However, the remaining variables were found to not be statistically significant because their p-values were $>$ 0.05.

Interpretation of parameter estimates

The adjusted odds ratios for the intercept and coefficients with a 95% confidence interval were as follows:

	O.R.	2.5 %	97.5 %
(Intercept)	2.9493	0.4520	21.8171
smoked1	2.8499	1.2318	6.8230
history.premature.labour1	2.6208	0.9436	7.8527
UTI1	2.9291	0.9589	9.8643
history.hypertension1	6.1299	1.3010	39.7462
mothers.weight.last.menstrual	0.9840	0.9685	0.9982

1. **Smoking during pregnancy**: Adjusting for all other factors, babies born to mothers who smoked during pregnancy are 3 times more likely to have a low birth weight than babies born to mothers who did not smoke during pregnancy.
2. **History of premature labour**: Adjusting for all other factors, babies born to mothers with a history of premature labour are 3 times more likely to have a low birth weight than babies born to mothers with no history of premature labour.
3. **Urinary Tract Infection**: Adjusting for all other factors, babies born to mothers having a urinary tract infection are 3 times more likely to have a low birth weight than babies born to mothers not having a urinary tract infection.
4. **History of hypertension**: Adjusting for all other factors, babies born to mothers with a history of hypertension are 6 times more likely to have a low birth weight than babies born to mothers with no history of hypertension.
5. **Mother's weight at their last menstrual**: For every unit increase in the mother's weight at their last menstrual holding all other variables constant, babies are 1.6% less likely to have a low birth weight.

Problem 2

Researchers in a certain county tracked flu cases requiring hospitalization in those residents aged 65 and older during a two-month period one winter.

They matched each case with 2 controls by sex and age (150 cases, 300 controls). They used medical records to determine whether cases and controls had received a flu vaccine shot and whether they had underlying lung disease.

Their interest was to determine if flu vaccination prevents hospitalization for flu (severe cases of flu). The underlying lung disease is a potential confounder. The dataset is **flumatch.csv** and variables are:

Variable	Code
Outcome	0 = Control 1 = Case
Vaccine	0 = not vaccinated 1 = vaccinated
Lung	0 = no underlying lung disease 1 = underlying lung disease
Id	Identifier for each matching group (1 case, 2 controls)

What conclusion do you think the researchers made?

Solution for Question 2

We are trying to see if flu infection among the elderly in wintertime has a relationship with flu vaccination and underlying lung disease.

Therefore, a binary regression model will be useful to help us achieve this goal.

Let $Y = outcome$ where if the individual is a patient, $Y = 1$ and if they are not a patient, $Y = 0$

Let $X_1 = vaccine$ (a vaccination for the flu) and $X_2 = lungdisease$ (underlying lung disease)

Stepwise regression based on the AIC was conducted, resulting in the following models:

```
## Start:  AIC=574.86
## flu.infection ~ 1
##
##                     Df Deviance    AIC
## + has.lung.disease  1   537.97 541.97
## + vaccinated        1   570.29 574.29
## <none>                  572.86 574.86
##
## Step:  AIC=541.97
## flu.infection ~ has.lung.disease
##
##                     Df Deviance    AIC
## + vaccinated        1   535.50 541.50
## <none>                  537.97 541.97
## - has.lung.disease  1   572.86 574.86
##
```

```
## Step:  AIC=541.5
## flu.infection ~ has.lung.disease + vaccinated
##
##                     Df Deviance    AIC
## <none>                   535.50 541.50
## - vaccinated         1   537.97 541.97
## - has.lung.disease   1   570.29 574.29
##
## Call:  glm(formula = flu.infection ~ has.lung.disease + vaccinated,
##     family = binomial)
##
## Coefficients:
##       (Intercept)  has.lung.disease1        vaccinated1
##           -0.9430             1.3379            -0.3453
##
## Degrees of Freedom: 449 Total (i.e. Null);  447 Residual
## Null Deviance:       572.9
## Residual Deviance: 535.5     AIC: 541.5
```

A more detailed description of the best model findings were as follows:

```
##
## Call:
## glm(formula = flu.infection ~ vaccinated + has.lung.disease,
##     family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3490  -0.8111  -0.6979   1.0150   1.7503
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.9430     0.1444  -6.532 6.47e-11 ***
## vaccinated1        -0.3453     0.2212  -1.561    0.118
## has.lung.disease1   1.3379     0.2292   5.837 5.33e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 572.86  on 449  degrees of freedom
## Residual deviance: 535.50  on 447  degrees of freedom
## AIC: 541.5
##
## Number of Fisher Scoring iterations: 4
```

Statistical significance of the model

The findings of the Likelihood Ratio Test were as follows:

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
449	572.8628	NA	NA	NA
447	535.4981	2	37.36469	0

Based on the results, the model is statistically significant because the p-value is « 0.05

Statistical significance of the coefficients

The vaccination variable does not appear to have a statistically significant relationship with the dependent variable because its p-value is > 0.05.

However, the underlying lung disease variable has a statistically significant relationship with the dependent variable because its p-value is « 0.05.

Interpretation of parameter estimates

The adjusted odds ratios for the intercept and coefficients with a 95% confidence interval were as follows:

	O.R.	2.5 %	97.5 %
(Intercept)	0.3895	0.2918	0.5143
vaccinated1	0.7080	0.4565	1.0879
has.lung.disease1	3.8110	2.4392	5.9993

1. **Flu vaccination**: Adjusting for all other factors, people who have received a flu vaccination are 61.05% less likely to get infected with the flu than people who are not vaccinated against the flu.
2. **Underlying lung disease**: Adjusting for all other factors, people who have underlying lung disease are 4 times more likely to get infected with the flu than people who are not vaccinated against the flu.

Conclusion

The relationship between flu infection and underlying lung disease should be studied in more depth.

Problem 3

The data set **serv.csv** gives part of data obtained during a 10 year follow up study on risk factors associated with death due to cancer for those serving in the military in Britain.

The number of deaths are recorded per person years for the pair combination of type of service(veteran or non veteran) and age category of the soldiers.

Are the two factors significantly associated with cancer deaths?

Solution for Question 3

When count data contains unequal time periods, the response variable should be the average number of events per unit time.

Therefore, the data around this question is as follows:

cancer	person.yrs	age	service	observed.rate	mean.cancer.per.100000.pyrs
6	60840	0-24	veteran	0.0000986	10
21	157175	25-29	veteran	0.0001336	13
54	176134	30-34	veteran	0.0003066	31
118	186514	35-39	veteran	0.0006327	63
97	135475	40-44	veteran	0.0007160	72
58	42620	45-49	veteran	0.0013609	136
56	25001	50-54	veteran	0.0022399	224
54	13710	55-59	veteran	0.0039387	394
34	6163	60-64	veteran	0.0055168	552
9	1575	65-69	veteran	0.0057143	571
2	273	70+	veteran	0.0073260	733
18	208487	0-24	non-veteran	0.0000863	9
60	303832	25-29	non-veteran	0.0001975	20
122	325421	30-34	non-veteran	0.0003749	37
191	312242	35-39	non-veteran	0.0006117	61
108	165597	40-44	non-veteran	0.0006522	65
74	54396	45-49	non-veteran	0.0013604	136
88	40716	50-54	non-veteran	0.0021613	216
120	33801	55-59	non-veteran	0.0035502	355
141	26618	60-64	non-veteran	0.0052972	530
108	17404	65-69	non-veteran	0.0062055	621
99	14146	70+	non-veteran	0.0069984	700

Because we are dealing with count data over a given period of time, we can apply the Poisson regression model provided that the mean and variance in the count data are as close as possible.

The mean and variance for our observed rates is as follows:

```
## [1] "Mean of the observed rates:"
```

```
## [1] 252.2273
```

```
## [1] "Variance of the observed rates:"
```

```
## [1] 64772.85
```

Because the mean and the variance differ significantly, a negative binomial regression model will be used as it is useful for count data that has a significant difference between its mean and its variance.

Stepwise regression based on the AIC was conducted, resulting in the following models:

```
## Start:  AIC=288.45
## mean.number.of.cancer.cases ~ 1
##
##             Df Deviance    AIC
## + age.group 10   0.0977 282.48
## <none>          26.0710 288.45
## + service    1  26.0697 290.45
##
## Step:  AIC=175.25
## mean.number.of.cancer.cases ~ age.group
##
##             Df Deviance    AIC
## <none>              8.0  175.3
## + service    1      7.5  176.8
## - age.group 10   5554.7 5702.0
```

```
##
## Call:  glm.nb(formula = mean.number.of.cancer.cases ~ age.group, init.theta = 3890101.522,
##     link = log)
##
## Coefficients:
##    (Intercept)  age.group25-29  age.group30-34  age.group35-39  age.group40-44
##         2.2513          0.5521          1.2751          1.8758          1.9755
## age.group45-49  age.group50-54  age.group55-59  age.group60-64  age.group65-69
##         2.6614          3.1423          3.6743          4.0421          4.1389
##   age.group70+
##         4.3231
##
## Degrees of Freedom: 21 Total (i.e. Null);  11 Residual
## Null Deviance:       5555
## Residual Deviance: 7.951     AIC: 177.3
```

A more detailed description of the best model findings were as follows:

```
##
## Call:
## glm.nb(formula = mean.number.of.cancer.cases ~ age.group, init.theta = 3890293.074,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0312  -0.4625   0.0000   0.4583   1.0169
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.2513     0.2294   9.813  < 2e-16 ***
## age.group25-29   0.5521     0.2880   1.917   0.0552 .
## age.group30-34   1.2751     0.2595   4.914 8.94e-07 ***
## age.group35-39   1.8758     0.2464   7.614 2.66e-14 ***
## age.group40-44   1.9755     0.2448   8.070 7.04e-16 ***
```

```
## age.group45-49   2.6614     0.2373  11.215  < 2e-16 ***
## age.group50-54   3.1423     0.2343  13.411  < 2e-16 ***
## age.group55-59   3.6743     0.2323  15.817  < 2e-16 ***
## age.group60-64   4.0421     0.2314  17.466  < 2e-16 ***
## age.group65-69   4.1389     0.2312  17.899  < 2e-16 ***
## age.group70+     4.3231     0.2309  18.720  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(3890293) family taken to be 1)
##
##     Null deviance: 5554.7051  on 21  degrees of freedom
## Residual deviance:    7.9507  on 11  degrees of freedom
## AIC: 177.25
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  3890293
##           Std. Err.:  81121968
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -153.253
```

Statistical significance of the model

A Likelihood Ratio test was conducted and the results were as follows:

Model	theta	Resid. df	2 x log-lik.	Test	df	LR stat.	Pr(Chi)
1	7.793561e-01	21	-286.4515		NA	NA	NA
age.group	3.890293e+06	11	-153.2529	1 vs 2	10	133.1987	0

Based on the results, the model is statistically significant because the p-value is « 0.05.

Statistical significance of the coefficients

Age: With the age group of 0-24 as the reference category, all dummy variables formed from the age variable except the 25-29 age group dummy variable were found to be statistically significant because their p-values were « 0.05.

Interpretation of parameter estimates

The adjusted odds ratio for the intercept and coefficients with a 95% confidence interval were as follows:

	O.R.	2.5 %	97.5 %
(Intercept)	9.5000	5.8429	14.4338
age.group25-29	1.7368	0.9982	3.1106
age.group30-34	3.5789	2.1987	6.1175
age.group35-39	6.5263	4.1325	10.9155

	O.R.	2.5 %	97.5 %
age.group40-44	7.2105	4.5821	12.0287
age.group45-49	14.3158	9.2551	23.5850
age.group50-54	23.1579	15.0733	37.9630
age.group55-59	39.4211	25.7763	64.4065
age.group60-64	56.9474	37.3113	92.9033
age.group65-69	62.7368	41.1217	102.3167
age.group70+	75.4211	49.4700	122.9403

Age

The findings for the different categories were as follows:

i) People whose ages fall in the 25-29 category were 73.68% more likely to contract cancer compared to those in the 0-24 age category.

ii) People whose ages fall in the 30-34 category were 4 times more likely to contract cancer compared to those in the 0-24 age category.

iii) People whose ages fall in the 35-39 category were 7 times more likely to contract cancer compared to those in the 0-24 age category.

iv) People whose ages fall in the 40-44 category were 7 times more likely to contract cancer compared to those in the 0-24 age category.

v) People whose ages fall in the 45-39 category were 14 times more likely to contract cancer compared to those in the 0-24 age category.

vi) People whose ages fall in the 50-54 category were 23 times more likely to contract cancer compared to those in the 0-24 age category.

vii) People whose ages fall in the 55-59 category were 39 times more likely to contract cancer compared to those in the 0-24 age category.

viii) People whose ages fall in the 60-64 category were 57 times more likely to contract cancer compared to those in the 0-24 age category.

ix) People whose ages fall in the 65-69 category were 63 times more likely to contract cancer compared to those in the 0-24 age category.

x) People whose ages fall in the 70+ category were 75 times more likely to contract cancer compared to those in the 0-24 age category.

Problem 4

The dataset **lungcancer.csv** has information of mortality by age and smoking status.

The dataset contains the following variables:

- Age at baseline: 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+
- Smoking status: 1 = never smoked, 2 = smoked cigars only, 3 = smoked cigarettes and cigars, 4 = smoked cigarettes only
- Population: number of male pensioners followed
- Deaths: number of deaths in a six-year period

Fit an appropriate model that describes mortality using age at baseline and smoking status.

Solution for Question 4

Let the response variable, $Y = mortality$, a variable consisting of 2 categories

If the subject is dead, $Y = 1$ and if the subject is living, $Y = 0$.

Let $X_1 = age$ (age at baseline status), a variable consisting of 9 categories

Let $X_2 = smoking\ status$, a variable consisting of 4 categories

Based on the available data, a binary logistic regression model can be used because the response variable has only two categories.

A binary logistic regression model can either be conditional or unconditional.

An unconditional regression model is expressed as

$$ln\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$ln\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\frac{P(Y=1)}{P(Y=0)} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}$$

A conditional model can be expressed as:

$$ln\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

$$ln\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

We first establish if there is a significant interaction, so we run stepwise regression using the binary conditional regression model as follows:

```
##
## Call:
## glm(formula = mortality ~ age + smoking.status + age:smoking.status,
##     family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.2770  -0.6344  -0.4820  -0.2586   2.6817
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -3.56797    0.23900 -14.929  < 2e-16 ***
## age45-49                   0.83893    0.32488   2.582  0.00981 **
## age50-54                   1.07433    0.33779   3.180  0.00147 **
## age55-59                   1.21746    0.27756   4.386 1.15e-05 ***
## age60-64                   1.47368    0.25831   5.705 1.16e-08 ***
## age65-69                   2.11484    0.25373   8.335  < 2e-16 ***
## age70-74                   2.56299    0.25447  10.072  < 2e-16 ***
## age75-79                   2.87066    0.26383  10.881  < 2e-16 ***
## age80+                     3.34811    0.26814  12.486  < 2e-16 ***
## smoking.status2            0.85254    0.41903   2.035  0.04190 *
## smoking.status3            0.18665    0.25310   0.737  0.46085
## smoking.status4            0.29082    0.25591   1.136  0.25578
## age45-49:smoking.status2 -0.36421    0.57849  -0.630  0.52897
## age50-54:smoking.status2 -0.06969    0.55793  -0.125  0.90060
## age55-59:smoking.status2 -0.67558    0.47414  -1.425  0.15420
## age60-64:smoking.status2 -0.62801    0.44204  -1.421  0.15540
## age65-69:smoking.status2 -0.90027    0.43579  -2.066  0.03884 *
## age70-74:smoking.status2 -0.90771    0.43540  -2.085  0.03709 *
## age75-79:smoking.status2 -0.71190    0.44107  -1.614  0.10652
## age80+:smoking.status2   -0.74826    0.44479  -1.682  0.09251 .
## age45-49:smoking.status3 -0.28664    0.34460  -0.832  0.40552
## age50-54:smoking.status3 -0.06756    0.35596  -0.190  0.84948
## age55-59:smoking.status3  0.19976    0.29318   0.681  0.49565
## age60-64:smoking.status3  0.28905    0.27360   1.056  0.29076
## age65-69:smoking.status3  0.07064    0.26975   0.262  0.79343
## age70-74:smoking.status3 -0.02172    0.27208  -0.080  0.93637
## age75-79:smoking.status3  0.05034    0.28528   0.176  0.85993
## age80+:smoking.status3    0.22510    0.30090   0.748  0.45441
## age45-49:smoking.status4 -0.26936    0.34862  -0.773  0.43973
## age50-54:smoking.status4  0.01695    0.35836   0.047  0.96228
## age55-59:smoking.status4  0.38037    0.29617   1.284  0.19903
## age60-64:smoking.status4  0.44950    0.27696   1.623  0.10460
## age65-69:smoking.status4  0.24052    0.27346   0.880  0.37911
## age70-74:smoking.status4  0.14532    0.27703   0.525  0.59989
## age75-79:smoking.status4  0.36978    0.29521   1.253  0.21036
## age80+:smoking.status4    0.16015    0.34079   0.470  0.63840
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49935  on 56121  degrees of freedom
## Residual deviance: 45066  on 56086  degrees of freedom
## AIC: 45138
##
## Number of Fisher Scoring iterations: 6

## Start:  AIC=49936.76
## mortality ~ 1
##
```

```
##                   Df Deviance   AIC
## + age              8    45313 45331
## + smoking.status  3    49736 49744
## <none>                  49935 49937
##
## Step:  AIC=45330.65
## mortality ~ age
##
##                   Df Deviance   AIC
## + smoking.status  3    45104 45128
## <none>                  45313 45331
## - age              8    49935 49937
##
## Step:  AIC=45127.66
## mortality ~ age + smoking.status
##
##                       Df Deviance   AIC
## <none>                      45104 45128
## + age:smoking.status 24    45066 45138
## - smoking.status      3    45313 45331
## - age                 8    49736 49744
```

```
##
## Call:  glm(formula = mortality ~ age + smoking.status, family = binomial)
##
## Coefficients:
##     (Intercept)         age45-49         age50-54         age55-59
##         -3.7106           0.5744           1.0477           1.4671
##        age60-64         age65-69         age70-74         age75-79
##          1.7934           2.2273           2.6018           3.0347
##          age80+  smoking.status2  smoking.status3  smoking.status4
##          3.5268           0.0805           0.2871           0.5412
##
## Degrees of Freedom: 56121 Total (i.e. Null);  56110 Residual
## Null Deviance:       49930
## Residual Deviance: 45100     AIC: 45130
```

Based on the above results, the interaction model had a slightly higher AIC than the best model, which turned out to be an unconditional binary regression model.

Furthermore, out of the 24 interaction dummy variables, only 2 were statistically significant because their p-values were < 0.05.

A more detailed description of the best model findings were as follows:

```
##
## Call:
## glm(formula = mortality ~ age + smoking.status, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3325  -0.6395  -0.5142  -0.2533   2.7330
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.71058    0.07142 -51.951  < 2e-16 ***
```

```
## age45-49         0.57444    0.08110   7.083 1.41e-12 ***
## age50-54         1.04774    0.07825  13.390  < 2e-16 ***
## age55-59         1.46706    0.06658  22.034  < 2e-16 ***
## age60-64         1.79340    0.06376  28.127  < 2e-16 ***
## age65-69         2.22728    0.06449  34.535  < 2e-16 ***
## age70-74         2.60176    0.06725  38.686  < 2e-16 ***
## age75-79         3.03467    0.07367  41.190  < 2e-16 ***
## age80+           3.52683    0.08380  42.088  < 2e-16 ***
## smoking.status2  0.08050    0.05452   1.476     0.14
## smoking.status3  0.28713    0.04424   6.490 8.58e-11 ***
## smoking.status4  0.54123    0.04581  11.814  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49935  on 56121  degrees of freedom
## Residual deviance: 45104  on 56110  degrees of freedom
## AIC: 45128
##
## Number of Fisher Scoring iterations: 6
```

Statistical significance of the binary logistic regression model

The findings of the Likelihood Ratio Test were as follows:

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
56121	49934.76	NA	NA	NA
56110	45103.66	11	4831.1	0

Based on the results, this model is statistically significant because the p-values of the Likelihood Ratio Test Statistic is « 0.05.

Statistical significance of coefficients

All coefficients associated with dummy variables derived from X_1 (with 40-44 age category as the reference), were statistically significant, with p-values « 0.05.

However, with the smoking status of 1 as the reference, only 2 of the 3 dummy variables derived from X_2 were statistically significant, with p-values « 0.05. the dummy variable for smoking status 2 (people who smoked cigars) was not statistically significant because its p-value was > 0.05.

Interpretation of parameter estimates

The adjusted odds ratio for the intercept and coefficients with a 95% confidence interval were as follows:

	O.R.	2.5 %	97.5 %
(Intercept)	0.0245	0.0212	0.0281
age45-49	1.7761	1.5154	2.0828

	O.R.	2.5 %	97.5 %
age50-54	2.8512	2.4470	3.3258
age55-59	4.3365	3.8115	4.9486
age60-64	6.0099	5.3127	6.8218
age65-69	9.2746	8.1867	10.5424
age70-74	13.4874	11.8393	15.4118
age75-79	20.7940	18.0197	24.0551
age80+	34.0160	28.8928	40.1306
smoking.status2	1.0838	0.9741	1.2062
smoking.status3	1.3326	1.2225	1.4541
smoking.status4	1.7181	1.5713	1.8804

1. **Age**

   The findings for the different categories were as follows:

   i) Adjusting for smoking status, people whose ages fall in the 45-49 category were 77.61% more likely to die compared to those in the 40-44 age category.

   ii) Adjusting for smoking status, people whose ages fall in the 50-54 category were 3 times more likely to die compared to those in the 40-44 age category.

   iii) Adjusting for smoking status, people whose ages fall in the 55-59 category were 4 times more likely to die compared to those in the 40-44 age category.

   iv) Adjusting for smoking status, people whose ages fall in the 60-64 category were 6 times more likely to die compared to those in the 40-44 age category.

   v) Adjusting for smoking status, people whose ages fall in the 65-69 category were 9 times more likely to die compared to those in the 40-44 age category.

   vi) Adjusting for smoking status, people whose ages fall in the 70-74 category were 13 times more likely to die compared to those in the 40-44 age category.

   vii) Adjusting for smoking status, people whose ages fall in the 75-79 category were 21 times more likely to die compared to those in the 40-44 age category.

   viii) Adjusting for smoking status, people whose ages fall in the 80+ category were 34 times more likely to die compared to those in the 40-44 age category.

2. **Smoking status**

   The findings for the different categories were as follows:

   i) Adjusting for age, people who smoked cigars only were 8.38% more likely to die compared to those who have never smoked.

   ii) Adjusting for age, people who smoked cigarettes and cigars were 33.26% more likely to die compared to those who have never smoked.

   iii) Adjusting for age, people who smoked cigarettes only were 71.81% more likely to die compared to those who have never smoked.

Problem 5

The data **mental.csv** comes from a study of mental health for a random sample of adult residents of Alachua County, Florida.

Mental impairment is ordinal, with categories (well, mild symptom formation, moderate symptom formation, impaired).

The study related mental impairment(mental) to two predictor variables;

- socioeconomic status (ses)(1=high, 0=low) and
- life events index (event) which is a composite measure of the number and severity of important life events such as birth of a child, a new job, divorce, or a death in family that occurred to the subject within the past three years.

(a) Fit a multinomial logistic model and interpret the results.
(b) Fit an ordinal logistic regression model and interpret the results.
(c) Predict the mental impairment of an individual in high socio economic status with a index of 8 using both model fits. Compare the results.

Solution for Question 5

(a) Multinomial logistic model

A multinomial logistic model is simply one where the response variable has m categories where $m > 2$, and one reference variable is used to build $m - 1$ number of binary logistic models within this one model.

Let

well=0 < *mild symptom formation*=1 < *moderate symptom formation*=2 < *impaired*=3

Let

$X_1 = SES$ (Socio-economic status) and $X_2 = events$ (the life events index)

$$ln\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$ln\left[\frac{P(Y=2)}{P(Y=0)}\right] = \beta_3 + \beta_4 X_1 + \beta_5 X_2$$

$$ln\left[\frac{P(Y=3)}{P(Y=0)}\right] = \beta_6 + \beta_7 X_1 + \beta_8 X_2$$

When SES is high, $X_1 = 1$ and when it is low, $X_1 = 0$

Stepwise regression using the multinomial logistic model was performed and the findings were as follows:

```
## # weights:  16 (9 variable)
## initial  value 55.451774
## iter  10 value 48.349823
## final  value 48.349131
## converged
```

```
## # weights:  8 (3 variable)
## initial  value 55.451774
## final  value 54.521026
## converged
## Start:  AIC=115.04
## mental.impairment ~ 1
##
## trying + socioeconomic.status
## # weights:  12 (6 variable)
## initial  value 55.451774
## iter  10 value 52.642356
## final  value 52.642355
## converged
## trying + life.events.index
## # weights:  12 (6 variable)
## initial  value 55.451774
## iter  10 value 50.908228
## final  value 50.908198
## converged
##                          Df      AIC
## + +life.events.index     6 113.8164
## <none>                   3 115.0421
## + +socioeconomic.status  6 117.2847
## # weights:  12 (6 variable)
## initial  value 55.451774
## iter  10 value 50.908228
## final  value 50.908198
## converged
##
## Step:  AIC=113.82
## mental.impairment ~ life.events.index
##
## trying - life.events.index
## # weights:  8 (3 variable)
## initial  value 55.451774
## final  value 54.521026
## converged
## trying + socioeconomic.status
## # weights:  16 (9 variable)
## initial  value 55.451774
## iter  10 value 48.349823
## final  value 48.349131
## converged
##                          Df      AIC
## <none>                    6 113.8164
## + +socioeconomic.status  9 114.6983
## - life.events.index      3 115.0421
## Call:
## multinom(formula = mental.impairment ~ life.events.index)
##
## Coefficients:
##   (Intercept) life.events.index
## 1  -0.7801769         0.2173599
```

```
## 2  -1.2512536         0.2011724
## 3  -2.4536330         0.4847313
##
## Residual Deviance: 101.8164
## AIC: 113.8164
```

A more detailed description of the best model findings were as follows:

```
## # weights:  12 (6 variable)
## initial  value 55.451774
## iter  10 value 50.908228
## final  value 50.908198
## converged
```

```
## Call:
## multinom(formula = mental.impairment ~ life.events.index, family = binomial)
##
## Coefficients:
##   (Intercept) life.events.index
## 1  -0.7801769         0.2173599
## 2  -1.2512536         0.2011724
## 3  -2.4536330         0.4847313
##
## Std. Errors:
##   (Intercept) life.events.index
## 1   0.7559415         0.1780997
## 2   0.8809727         0.2019959
## 3   1.0490171         0.2014681
##
## Residual Deviance: 101.8164
## AIC: 113.8164
```

Statistical significance of the model:

A likelihood ratio test was conducted and the results were as follows:

Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
1	117	109.0421		NA	NA	NA
life.events.index	114	101.8164	1 vs 2	3	7.225656	0.0650428

The p-value is > 0.05 $\therefore$ the model is not statistically significant.

Statistical significance of the coefficient

Comparing mild mental illness to mental wellness: *Life events*

$H_0 : \beta_{events} = 0$ vs. $H_1 : \beta_{events} \neq 0$

Let $\alpha = 0.05$ *be the level of significance*

Let z be the test statistic

$$z_{cal} = \frac{\beta_{events}}{s.e.\beta_{events}} = \frac{0.2174}{0.1781} = 1.2207$$

$$z_{tab} = z_{0.975} = 1.96$$

$z_{cal} < z_{tab}$ $\therefore$ fail to reject H_0; the statistical association between life events and the dependent variable is not significant.

Comparing moderate mental illness to mental wellness: *Life events*

$H_0 : \beta_{events} = 0$ vs. $H_1 : \beta_{events} \neq 0$

Let $\alpha = 0.05$ *be the level of significance*

Let z be the test statistic

$$z_{cal} = \frac{\beta_{events}}{s.e.\beta_{events}} = \frac{0.2012}{0.2020} = 0.9960$$

$$z_{tab} = z_{0.975} = 1.96$$

$z_{cal} < z_{tab}$ $\therefore$ fail to reject H_0; the statistical association between life events and the dependent variable holding all other variables constant is not significant.

Comparing mental impairment to mental wellness: *Life events*

$H_0 : \beta_{events} = 0$ vs. $H_1 : \beta_{events} \neq 0$

Let $\alpha = 0.05$ *be the level of significance*

Let z be the test statistic

$$z_{cal} = \frac{\beta_{events}}{s.e.\beta_{events}} = \frac{0.4847}{0.2015} = 2.4055$$

$$z_{tab} = z_{0.975} = 1.96$$

$z_{cal} > z_{tab}$ $\therefore$ reject H_0; the statistical association between life events and the dependent variable is significant.

Interpretation of parameter estimates:

Comparing mild mental illness to mental wellness: *Life events*

$$O.R._{events} = e^{0.2174} = 1.2428$$

For every unit increase in life events a person is 24.28% more likely to have mild mental illness.

Comparing moderate mental illness to mental wellness: *Life events*

$$O.R._{events} = e^{0.2012} = 1.2229$$

For every unit increase in life events a person is 22.29% more likely to have moderate mental illness.

Comparing mental impairment to mental wellness: *Life events*

$$O.R._{events} = e^{0.4847} = 1.6237$$

For every unit increase in life events, holding socioeconomic status constant, a person is 62.37% more likely to have moderate mental illness.

(b) Ordinal logistic model

An ordinal logistic model is where the event under study is a category or set of lower-ranked categories

Let

well=0 < *mild symptom formation*=1 < *moderate symptom formation*=2 < *impaired*=3

$$\frac{P(Y \leq 0)}{P(Y > 0)} = \frac{P(Y = 0)}{P(Y = 1) \; or \; P(Y = 2) \; or \; P(Y = 3)}$$

$$\frac{P(Y \leq 1)}{P(Y > 1)} = \frac{P(Y = 0) \; or \; P(Y = 1)}{P(Y = 2) \; or \; P(Y = 3)}$$

$$\frac{P(Y \leq 2)}{P(Y > 2)} = \frac{P(Y = 0) \; or \; P(Y = 1) \; or \; P(Y = 2)}{(Y = 3)}$$

Let

$X_1 = SES$ (Socio-economic status) and $X_2 = events$ (the life events index)

$$ln\left[\frac{P(Y \leq 0)}{P(Y > 0)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$ln\left[\frac{P(Y \leq 1)}{P(Y > 1)}\right] = \beta_3 + \beta_1 X_1 + \beta_2 X_2$$

$$ln\left[\frac{P(Y \leq 2)}{P(Y > 2)}\right] = \beta_4 + \beta_1 X_1 + \beta_2 X_2$$

When SES is high, $X_1 = 1$ and when it is low, $X_1 = 0$

Stepwise regression using the ordinal logistic model was performed and the findings were as follows:

```
## Start:  AIC=115.04
## mental.impairment ~ 1
##
##                         Df    AIC
## + life.events.index      1 110.53
## + socioeconomic.status   1 114.87
## <none>                     115.04
##
## Step:  AIC=110.53
## mental.impairment ~ life.events.index
##
##                         Df    AIC
## + socioeconomic.status   1 109.10
## <none>                     110.53
## - life.events.index      1 115.04
##
## Step:  AIC=109.1
## mental.impairment ~ life.events.index + socioeconomic.status
##
##                         Df    AIC
## <none>                     109.10
## - socioeconomic.status   1 110.53
## - life.events.index      1 114.87
```

```
## Call:
## polr(formula = mental.impairment ~ life.events.index + socioeconomic.status)
##
## Coefficients:
##     life.events.index socioeconomic.status1
##             0.3188613            -1.1112310
##
## Intercepts:
##        0|1        1|2        2|3
## -0.2819031  1.2127926  2.2093721
##
## Residual Deviance: 99.0979
## AIC: 109.0979
```

Therefore, the best model was as follows:

```
## Call:
## polr(formula = mental.impairment ~ life.events.index + socioeconomic.status)
##
## Coefficients:
##                          Value Std. Error t value
## life.events.index       0.3189     0.1210   2.635
## socioeconomic.status1 -1.1112     0.6109  -1.819
##
## Intercepts:
##     Value   Std. Error t value
## 0|1 -0.2819  0.6423    -0.4389
## 1|2  1.2128  0.6607     1.8357
## 2|3  2.2094  0.7210     3.0644
##
```

```
## Residual Deviance: 99.0979
## AIC: 109.0979
```

Statistical significance of the model:

A likelihood ratio test was conducted and the results were as follows:

Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
1	37	109.0421		NA	NA	NA
life.events.index + socioeconomic.status	35	99.0979	1 vs 2	2	9.944157	0.0069287

The p-value is < 0.05 $\therefore$ the model is statistically significant.

Statistical significance of coefficients:

1. **Life events**

   $H_0 : \beta_{events} = 0$ vs. $H_1 : \beta_{events} \neq 0$

   Let $\alpha = 0.05$ *be the level of significance*

   Let z be the test statistic

   $$z_{cal} = \frac{\beta_{events}}{s.e.\beta_{events}} = \frac{0.3189}{0.1210} = 2.6355$$

   $$z_{tab} = z_{0.975} = 1.96$$

   $z_{cal} > z_{tab}$ $\therefore$ reject H_0; the statistical association between life events and the dependent variable variable holding all other variables constant is significant.

2. **SES**

   $H_0 : \beta_{SES} = 0$ vs. $H_1 : \beta_{SES} \neq 0$

   Let $\alpha = 0.05$ *be the level of significance*

   Let z be the test statistic

   $$z_{cal} = \frac{\beta_{SES}}{s.e.\beta_{SES}} = \frac{-1.1112}{0.6109} = -1.8190$$

   $$z_{tab} = z_{0.025} = -1.96$$

   $z_{cal} > z_{tab}$ $\therefore$ fail to reject H_0; the statistical association between SES and the dependent variable variable holding all other variables constant is not significant.

Interpretation of parameter estimates:

1. **Life events**

$$O.R._{events} = e^{0.3189} = 1.3756$$

For every unit increase in life events, holding socioeconomic status constant, a person is 37.56% more likely to have mild mental illness symptoms or to have moderate mental illness symptoms or to be mentally impaired (with having mental wellness as the reference category).

2. **SES**

$$O.R._{SES} = e^{-1.1112} = 0.3292$$

A person with a high socioeconomic status, holding life events constant, is 67.08% less likely to have mild mental illness symptoms or to have moderate mental illness symptoms or to be mentally impaired (with having mental wellness as the reference category) compared to people of low socioeconomic status.

(c) Predicting the mental impairment of an individual

Prediction of mental impairment using the multinomial model

$$\frac{P(Y=1)}{P(Y=0)} = e^{\beta_0+\beta_1 X_2} = e^{-0.7802-0.2174X_2}$$

$$\frac{P(Y=2)}{P(Y=0)} = e^{\beta_2+\beta_3 X_2} = e^{-1.2513+0.2012X_2}$$

$$\frac{P(Y=3)}{P(Y=0)} = e^{\beta_4+\beta_5 X_2} = e^{-2.4536+0.4847X_2}$$

If $X_1 = 1$ and $X_2 = 8$,

$$\frac{P(Y=1)}{P(Y=0)} = e^{-0.7802-0.2174(8)} = e^{-2.5194} = 0.0805$$

$$\frac{P(Y=2)}{P(Y=0)} = e^{-1.2513+0.2012(8)} = e^{0.3583} = 1.4309$$

$$\frac{P(Y=3)}{P(Y=0)} = e^{-2.4536+0.4847(8)} = e^{1.4240} = 4.1537$$

i)

$$\frac{P(Y=1)}{P(Y=0)} = 0.0805$$

$$P(Y=0) \times \frac{P(Y=1)}{P(Y=0)} = 0.0805[P(Y=0)]$$

$$P(Y=1) = 0.0805[P(Y=0)]$$

ii)

$$\frac{P(Y=2)}{P(Y=0)} = 1.4309$$

$$P(Y=0) \times \frac{P(Y=2)}{P(Y=0)} = 1.4309[P(Y=0)]$$

$$P(Y=2) = 1.4309[P(Y=0)]$$

iii)

$$\frac{P(Y=3)}{P(Y=0)} = 4.1537$$

$$P(Y=0) \times \frac{P(Y=3)}{P(Y=0)} = 4.1537[P(Y=0)]$$

$$P(Y=3) = 4.1537[P(Y=0)]$$

iv)

$$P(Y=0) + P(Y=1) + P(Y=2) + P(Y=3) = 1$$

$$P(Y=0) + 0.0805[P(Y=0)] + 1.4309[P(Y=0)] + 4.1537[P(Y=0)] = 1$$

$$[1 + 0.0805 + 1.4309 + 4.1537]P(Y=0) = 1$$

$$\frac{6.6651[P(Y=0)]}{6.6651} = \frac{1}{6.6651}$$

$$P(Y=0) = 0.1500$$

$$\therefore$$

$$P(Y=1) = 0.0805[P(Y=0)] = 0.0805 \times 0.1500 = 0.0121$$

$$P(Y=2) = 1.4309[P(Y=0)] = 1.4309 \times 0.1500 = 0.2146$$

$$P(Y=3) = 4.1537[P(Y=0)] = 4.1537 \times 0.1500 = 0.6231$$

Based on the results, an individual with a life events index of 8 is most likely to become mentally impaired; this outcome had the highest probability of 0.6231.

Prediction of mental impairment using the ordinal logistic model

$$\frac{P(Y \leq 0)}{P(Y > 0)} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2} = e^{-0.2819 - 1.1112 X_1 + 0.3189 X_2}$$

$$\frac{P(Y \leq 1)}{P(Y > 1)} = e^{\beta_3 + \beta_1 X_1 + \beta_2 X_2} = e^{1.2128 - 1.1112 X_1 + 0.3189 X_2}$$

$$\frac{P(Y \leq 2)}{P(Y > 2)} = e^{\beta_4 + \beta_1 X_1 + \beta_2 X_2} = e^{2.2094 - 1.1112 X_1 + 0.3189 X_2}$$

If $X_1 = 1$ and $X_2 = 8$,

$$\frac{P(Y \leq 0)}{P(Y > 0)} = e^{-0.2819 - 1.1112(1) + 0.3189(8)} = e^{1.1581} = 3.1839$$

$$\frac{P(Y \leq 1)}{P(Y > 1)} = e^{1.2128 - 1.1112(1) + 0.3189(8)} = e^{2.6528} = 14.1937$$

$$\frac{P(Y \leq 2)}{P(Y > 2)} = e^{2.2094 - 1.1112(1) + 0.3189(8)} = e^{3.6494} = 38.4516$$

i)

$$P(Y=0) + P(Y=1) + P(Y=2) + P(Y=3) = 1$$

$$P(Y=0) = 1 - P(Y=1) - P(Y=2) - P(Y=3)$$

ii)

$$\frac{P(Y \leq 0)}{P(Y > 0)} = 3.1839$$

$$\frac{P(Y=0)}{P(Y=1) + P(Y=2) + P(Y=3)} = 3.1839$$

$$[P(Y=1)+P(Y=2)+P(Y=3)] \times \frac{P(Y=0)}{P(Y=1) + P(Y=2) + P(Y=3)} = 3.1839[P(Y=1)+P(Y=2)+P(Y=3)]$$

$$P(Y=0) = 3.1839[P(Y=1) + P(Y=2) + P(Y=3)]$$

$$1 - P(Y=1) - P(Y=2) - P(Y=3) = 3.1839[P(Y=1)] + 3.1839[P(Y=2)] + 3.1839[P(Y=3)]$$

$$1 = [3.1839+1]P(Y=1) + [3.1839+1]P(Y=2) + [3.1839+1]P(Y=3)$$

$$\frac{1}{4.1839} = \frac{4.1839[P(Y=1) + P(Y=2) + P(Y=3)]}{4.1839}$$

$$P(Y=1) + P(Y=2) + P(Y=3) = 0.2390$$

$$P(Y=1) = 0.2390 - P(Y=2) - P(Y=3)$$

iii)

$$\frac{P(Y \leq 1)}{P(Y > 1)} = 14.1937$$

$$\frac{P(Y=0) + P(Y=1)}{P(Y=2) + P(Y=3)} = 14.1937$$

$$\frac{[1 - P(Y=1) - P(Y=2) - P(Y=3)] + P(Y=1)}{P(Y=2) + P(Y=3)} = 14.1937$$

$$\frac{1 + [1-1]P(Y=1) - P(Y=2) - P(Y=3)}{P(Y=2) + P(Y=3)} = 14.1937$$

$$\frac{1 - P(Y=2) - P(Y=3)}{P(Y=2) + P(Y=3)} = 14.1937$$

$$[P(Y=2) + P(Y=3)] \times \frac{1 - P(Y=2) - P(Y=3)}{P(Y=2) + P(Y=3)} = 14.1937[P(Y=2) + P(Y=3)]$$

$$1 - P(Y=2) - P(Y=3) = 14.1937[P(Y=2)] + 14.1937[P(Y=3)]$$

$$1 = [14.1937+1]P(Y=2) + [14.1937+1]P(Y=3)$$

$$\frac{1}{15.1937} = \frac{15.1937[P(Y=2) + P(Y=3)]}{15.1937}$$

$$P(Y=2) + P(Y=3) = 0.0658$$

$$P(Y=2) = 0.0658 + P(Y=3)$$

iv)

$$\frac{P(Y \leq 2)}{P(Y > 2)} = 38.4516$$

$$\frac{P(Y = 0) + P(Y = 1) + P(Y = 2)}{P(Y = 3)} = 38.4516$$

$$\frac{[1 - P(Y = 1) - P(Y = 2) - P(Y = 3)] + P(Y = 1) + P(Y = 2)}{P(Y = 3)} = 38.4516$$

$$\frac{1 + [1 - 1]P(Y = 1) + [1 - 1]P(Y = 2) - P(Y = 3)}{P(Y = 3)} = 38.4516$$

$$\frac{1 - P(Y = 3)}{P(Y = 3)} = 38.4516$$

$$P(Y = 3) \times \frac{1 - P(Y = 3)}{P(Y = 3)} = 38.4516[P(Y = 3)]$$

$$1 - P(Y = 3) = 38.4516[P(Y = 3)]$$

$$1 = [38.4516 + 1]P(Y = 3)$$

$$\frac{1}{39.4516} = \frac{39.4516[P(Y = 3)]}{39.4516}$$

$$P(Y = 3) = 0.0253$$

$$\therefore$$

$$P(Y = 2) = 0.0658 + P(Y = 3) = 0.0658 + 0.0253 = 0.0911$$

$$P(Y = 1) = 0.2390 - P(Y = 2) - P(Y = 3) = 0.2390 - 0.0911 - 0.0253 = 0.1226$$

$$P(Y = 0) = 1 - P(Y = 1) - P(Y = 2) - P(Y = 3) = 1 - 0.1226 - 0.0911 - 0.0253 = 0.7610$$

Based on the results, an individual of a high socio-economic status with a life events index of 8 is most likely to have mental wellness; this outcome had the highest probability of 0.7610.

Comparison of both predictions

The predictions from both models were on complete ends of the spectrum; the multinomial model predicted that the individual under study would be mentally impaired while the ordinal logistic model prediction was that the individual was mentally well. It would therefore indicate that choice of model to be used for analysis is critical.

Problem 6

A clinical trial for the treatment of small-cell lung cancer was carried out where patients were randomly assigned to two treatment groups:

- Sequential therapy (the same combination of chemotherapeutic agents administered in each treatment cycle)
- Alternating therapy (three different combinations alternated from cycle to cycle)

Gender was considered a potential effect modifier.

The results of the trial were as follows:

Response to therapy was categorized as either

- Progressive disease,
- No change,
- Partial remission or
- Complete remission

Treatment therapy	Gender	Progressive Disease	No Change	Partial Remission	Complete Remission
Sequential	Male	68	51	39	36
	Female	14	27	12	9
Alternating	Male	83	87	56	60
	Female	32	17	12	7

Fit a proportional odds logit model and interpret the estimated treatment effect.

Solution for Question 6

A proportional odds logit model is one where the intercept is dependent on the category level.

When we observe the categories of the dependent variable (Response to therapy), *Progressive disease* should be the lowest rank (1) because it indicates that the drug under study is worsening the patient's disease and *Complete remission* should be the highest rank (4) because it indicates that the drug completely cured the patient - it is the result that we would like to see most.

Let

Progressive disease=1 < *No change*=2 < *Partial remission*=3 < *Complete remission*=4

$$\frac{P(Y \leq 1)}{P(Y > 1)} = \frac{P(Y = 1)}{P(Y = 2) \text{ or } P(Y = 3) \text{ or } P(Y = 4)}$$

$$\frac{P(Y \leq 2)}{P(Y > 2)} = \frac{P(Y = 1) \text{ or } P(Y = 2)}{P(Y = 3) \text{ or } P(Y = 4)}$$

$$\frac{P(Y \leq 3)}{P(Y > 3)} = \frac{P(Y = 1) \text{ or } P(Y = 2) \text{ or } P(Y = 3)}{(Y = 4)}$$

Let

$X_1 = Therapy$ and $X_2 = Gender$

$$ln\left[\frac{P(Y \leq 1)}{P(Y > 1)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$ln\left[\frac{P(Y \leq 2)}{P(Y > 2)}\right] = \beta_3 + \beta_1 X_1 + \beta_2 X_2$$

$$ln\left[\frac{P(Y \leq 3)}{P(Y > 3)}\right] = \beta_4 + \beta_1 X_1 + \beta_2 X_2$$

When treatment is sequential, $X_1 = 1$ and when it is alternating, $X_1 = 0$

When the patient is male, $X_2 = 1$ and when they are female, $X_2 = 0$

Once this data was used to fit a proportional odds logit model, the findings were as follows:

```
## Start:  AIC=1660.2
## ResponseToTherapy ~ 1
##
##           Df    AIC
## + Gender   1 1659.1
## <none>       1660.2
## + Therapy  1 1662.2
##
## Step:  AIC=1659.09
## ResponseToTherapy ~ Gender
##
##           Df    AIC
## <none>       1659.1
## - Gender   1 1660.2
## + Therapy  1 1661.1
```

```
## Call:
## polr(formula = ResponseToTherapy ~ Gender)
##
## Coefficients:
##    Gender
## 0.3108506
##
## Intercepts:
##        1|2        2|3        3|4
## -0.5009262  0.7385563  1.7398409
##
## Residual Deviance: 1651.093
## AIC: 1659.093
```

Therefore, the best model was as follows:

```
## Call:
## polr(formula = ResponseToTherapy ~ Gender)
##
## Coefficients:
##         Value Std. Error t value
## Gender 0.3109     0.1767   1.759
##
```

```
## Intercepts:
##     Value   Std. Error t value
## 1|2 -0.5009  0.1610    -3.1104
## 2|3  0.7386  0.1622     4.5530
## 3|4  1.7398  0.1763     9.8670
##
## Residual Deviance: 1651.093
## AIC: 1659.093
```

Statistical significance of the model

A likelihood ratio test was conducted and the results were as follows:

Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
1	607	1654.204		NA	NA	NA
Gender	606	1651.093	1 vs 2	1	3.111229	0.0777543

The p-value is > 0.05 $\therefore$ the model is not statistically significant.

Statistical significance of the coefficient: *Gender*

$H_0 : \beta_{gender} = 0$ vs. $H_1 : \beta_{gender} \neq 0$

Let $\alpha = 0.05$ *be the level of significance*

Let z be the test statistic

$$z_{cal} = \frac{\beta_{gender}}{s.e.\beta_{gender}} = \frac{0.3109}{0.1767} = 1.7595$$

$$z_{tab} = z_{0.975} = 1.96$$

$z_{cal} < z_{tab}$ $\therefore$ fail to reject H_0; the statistical association between gender and the dependent variable variable holding all other variables constant is not significant.

Interpretation of parameter estimate: *Gender*

$$O.R._{gender} = e^{0.3109} = 1.3647$$

A male patient is 36.47% more likely to have progressive disease or no change in illness or partial remission (with the complete remission as the reference category) compared to a female patient.

Problem 7

Critique the two journal articles on physical inactivity and adolescent risk behaviour.

Solution for Question 7

Physical Activity Article Critique

Toriola (2018) in their article A Multinomial Logistic Regression Analysis of the Association Between Physical Activity and Body Weight Status of University Women in Riyadh, Saudi Arabia sought to study the relationship between physical activity level and body weight among the population of university-going women in the kingdom of Saudi Arabia with a view of addressing growing obesity in Saudi Arabia and the Middle East region in general. While the researcher is solving a real-world problem and uses best practice in research study design, data collection and analysis, model comparisons, deeper interpretation of results and findings that take into account more variables were needed to make a stronger model.

Introduction

The researcher introduces their article by stating that past research indicated that "habitual physical activity lowers CVD [cardiovascular disease] risk and the outcome has been assumed to be facilitated by healthy body weight". They then further elaborate that people have fallen short of recommended levels of physical activity and body weight at normal values. Furthermore, more and more Saudi women according to the researcher have been found to develop cardiometabolic diseases which have obesity as a research.

The researcher's research is indeed relevant due to more countries experiencing the obesity of their citizens, particularly children, as a challenge (Onis et. al., 2010). The researcher also highlights research from reputable global bodies such as the World Health Organization to make his case.

While the researcher touches on obesity, they do not take into account nutrition which is a key determinant (Bopkin, 2001; Swinburn et. al., 2004); they focus more on physical exercise. It would be helpful to also consider influences outside an individual (Spence and Lee, 2002) such as cultural norms and how they influence physical activity levels, particularly in the Saudi community. These could be a contributing factor to physical inactivity.

Research Methodology

The researcher selected a cross-sectional survey design, with physical activity levels and Body Mass Index (BMI) characteristics as their variables of interest in the study and a large sample size (n>30) was taken resulting in 573 observations.

Measures of central tendency were used to evaluate the respondents age, weight, height, BMI and METs score. METs scores, in particular, were stratified into three physical activity level categories before measures of central tendency were taken: Low, where METs < 500; Moderate, where $500 \leq$ METs ≤ 1499; Vigorous, where METs > 500. Finally BMI data was put into four categories: Underweight (11.7%), Normal weight (65.1%), Overweight (18.7%) and Obese (4.5%).

One-way ANOVA was conducted on the physical activity level categories with age, body weight, height and BMI as treatments. This was then followed by a multinomial logistic regression model of BMI and physical activity level as categorical variables with BMI as the predictor variable and physical activity level as the dependent variable.

All statistical tests were two-tailed and the 95% confidence level was selected.

It is commendable that the researcher observed global best practice in getting the respondent's consent, taking the height and weight of respondents and measuring their physical activity levels. It is also commendable that a pilot test was conducted to ensure reliable data collection and current best practice for data analysis like use of statistical software like SPSS was observed.

However, using IPAQ as the entire questionnaire limited the study because it is a global standardized questionnaire that may not adequately account for cultural norms. A longitudinal study design would have been beneficial compared to a cross-sectional study to observe additional variables outside of the individual (Spence and Lee, 2002) and minimize the self-reporting aspect that is in the existing study. The categorical variables used in the researcher's model could also be interpreted as ordinal in nature; therefore an ordinal logistic regression model could have been used and the two models could have been compared in order to select the better model.

Results & Discussion

The researcher stressed that "participation in moderate to vigorous PA [physical activity level] irrespective of an individual's body weight classification is beneficial to health". The researcher also appreciates that gender, race and ethnicity also affect BMI and has called for further studies that can generate more generalizable findings and have larger sample sizes.

The researcher could have included the statistical significance of the entire multinomial logistic regression model in order for the reader to have a frame of reference for further interpretations of the model. Furthermore, transforming the reported Odds Ratios into percentages would ease interpretation surrounding the likelihoods of being underweight, overweight or obese. Interpretations of likelihood also need to ensure that the results are reported as being compared to the reference category, which in this study was normal weight.

Concluding Remark

This article is addressing a real-world problem and can open the door for further research into weight problems facing the Middle East.

References

Toriola, O., 2018. A multinomial logistic regression analysis of the association between physical activity and body weight status of university women in Riyadh, Saudi Arabia. *Asian Journal of Scientific Research*, 11, pp.145-150.

De Onis, M., Blössner, M. and Borghi, E., 2010. Global prevalence and trends of overweight and obesity among preschool children. *The American journal of clinical nutrition*, 92(5), pp.1257-1264.

Popkin, B.M., 2001. The nutrition transition and obesity in the developing world. *The Journal of Nutrition*, 131(3), pp.871S-873S.

Swinburn, B.A., Caterson, I., Seidell, J.C. and James, W.P.T., 2004. Diet, nutrition and the prevention of excess weight gain and obesity. *Public Health Nutrition*, 7(1a), pp.123-146.

Spence, J.C. and Lee, R.E., 2003. Toward a comprehensive model of physical activity. *Psychology of sport and exercise*, 4(1), pp.7-24.

Adolescent Risk Behaviour Article Critique

Peng and Nichols (2003) in their article Using multinomial logistic models to predict adolescent behavioral risk sought to apply multinomial logistic regression models on existing data on a sample of 432 adolescents that were in junior high (Grade 7-9) in 1988. This research was done in order to help psychologists and educators identify adolescents at the greatest behavioral risk and help them differentiate between categories of behavioural risk among adolescents. This differentiation could then help build a profile that can help personalize the prevention programs that psychologists and educators currently offer.

While this article is solving a real-world problem and did statistical analysis best practice such as conducting cross-validation and having a comparison of estimated probabilities with actual categorizations, the authors could consider enriching the article with more past studies that tie their identified covariates to the dependent variable e.g. briefly citing works where psychologists discussed the theoretical link between behavioural risk in adolescents with gender so that they can strengthen their case for selection of that variable. They could have also explored more than one model. For instance, they could have made the FAMILY variable continuous in one instance and categorical in another and evaluated the two; they could have specified an unconditional multinomial regression model where the dependent variable is not an ordered categorical variable and compared it to a different model where the dependent variable is an ordered categorical variable.

Introduction

The data source selected for the study was given by one of the authors of a 1993 article, G. M. Ingersoll was gathered from 517 junior high school students in 1988; ; upon realizing that there was missing data, those observations were omitted such that the final sample size became 432.

Each student filled two data collection tools to generate a self report of their risky behaviour: Rosenberg's Self Esteem Inventory and the Health Behaviour Questionnaire.

Consequently, the researchers declared their research hypothesis as "*the likelihood that an adolescent is at high, medium, or low behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk, and self esteem.*"

Bearing this in mind, the researchers then proceeded to convert the continuous behavioral risk score in the original data to a categorical variable by applying statistical rationale; the low/medium behavioural risk boundary was selected to be the median of the continuous variable and the medium/high behavioural risk boundary was selected to be the sum of the mean and 1 standard deviation of the original variable.

Measures of central tendency were applied to the rest of the variables in the dataset, stratifying the remaining categorical variables by the newly formed dependent variable - the categorical behavioural risk variable (hereafter known as the RISK variable). Most boys when classified fell in the high and medium categories and most children raised in single-parent homes fell in the medium category. This made the gender of the respondent (hereafter referred to as GENDER) and family structure (hereafter referred to as FAMILY) variables worth including in the model. A significant percentage (12.27%) of the respondents indicated that they would like to drop out of school; therefore, the intention of the child to dropout (hereafter referred to as DROPOUT) was included in the model. Finally, their emotional risk score and the self-esteem score of the child (hereafter referred to as EMOTION and ESTEEM respectively) were included in the model.

The multinomial logistic regression model was selected by the researchers because it effectively and reliably demonstrated how the self-reported characteristics of adolescents can impact their behavioural risk level and quantifies their net effect on each dependent variable category relative to the reference category.

The researchers clearly explained the context around data collection which increased the validity of their real-world data set. However, to increase the credibility of their selection criteria for the ESTEEM and EMOTION variables, it would be advisable to include past studies that theoretically link each of them to the dependent variable.

The researchers appear to have selected the ordinal form of the complex multinomial logistic model. It would have been interesting to evaluate a form of this model where the dependent variable was an unordered categorical variable or where FAMILY was a categorical variable instead of a continuous one. Comparing more than one model rather than only using the null model for comparison with the existing model would have allowed further discussion of the different types of complex multinomial models and would have allowed the researchers to select the best model.

The Multinomial Logistic Model on real-life data

Let $Y = 1$ *or* 2 *or* 3 where low behavioural risk (3) is the reference category; this is an ordered categorical variable

p_1 = probability of high behavioural risk = $P(Y = 1)$

p_2 = probability of medium behavioural risk = $P(Y = 2)$

The log of odds of Y is also called the Logit of Y

$$Logit(p_1) = natural\ log\ (odds) = ln$$

$$Logit(p_1 + p_2) = natural\ log\ (odds) = ln$$

The model constraint: Spi = 1

The constants in the model are first Y intercept, α_1, the second Y intercept, α_2 and the gradient of the resulting sigmoid curves, b. All three constants are determined by the maximum likelihood method.

Let

$X_1 = GENDER$ where $X_1 = 1$ if a boy and $X_1 = 0$ if a girl; this is a categorical variable

$X_2 = DROPOUT$ where $X_2 = 1$ if the response is 'Yes, I intend to drop out' and $X_2 = 0$ if the response is 'No, I don't intend to drop out'; this is a categorical variable

$X_3 = FAMILY$ where $X_3 = 1$ if from an intact family, $X_3 = 2$ if from a family with 1 step-parent and $X_3 = 3$ if from a family with a single-parent family; this is a continuous variable

$X_4 = EMOTION$; this is a continuous variable

$X_5 = ESTEEM$; this is a continuous variable

Let $e = 2.71828$, the base of the system of natural logarithms

Regarding the individual covariate coefficients:

- There are two hypotheses to be evaluated

  $H_0 : \beta_j = 0$ i.e. each and every one of the coefficients in the model are not statistically significant

  $H_1 : \beta_j \neq 0$ i.e. at least one of the coefficients in the model is statistically significant
- The odds ratio e^{β_j} is the change in the odds of Y given a unit change in X_j

Therefore, the resulting model will be:

$$p_1 = \frac{e^{\alpha_1+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5}}{1 - e^{\alpha_1+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5}}$$

$$p_1 + p_2 = \frac{e^{\alpha_2+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5}}{1 - e^{\alpha_2+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5}}$$

$$p_3 = 1 - (p_1 + p_2)$$

Model Results

The model defined above was fitted onto SAS PROC LOGISTIC (Version 8e, SAS Institute Inc., 1999), statistical computer software, and the resulting model was as follows:

$$Logit(\hat{p}_1) = -0.6211 + 1.1070X_1 + 2.1818X_2 + 0.4135X_3 + 0.00738X_4 - 0.0488X_5$$

$$Logit(\hat{p}_1 + \hat{p}_2) = 2.5220 + 1.1070X_1 + 2.1818X_2 + 0.4135X_3 + 0.00738X_4 - 0.0488X_5$$

The adjusted O.Rs for each of the covariates were obtained as shown:

Variable	Variable Name	Coefficient(β_j)	Adjusted Odds Ratio (e^{β_j})
X_1	GENDER	1.1070	3.0253
X_2	DROPOUT	2.1818	8.8622
X_3	FAMILY	0.4135	1.5121
X_4	EMOTION	0.00738	1.0074
X_5	ESTEEM	0.0488	1.0500

Based on the model:

- Boys are 3 times more likely to have high or medium behavioural risk compared to girls, holding all other covariates constant.
- Adolescents that intend to drop out of school are 9 times more likely to have high or medium behavioural risk compared to adolescents that don't intend to do so, holding all other covariates constant.
- For every unit increase in family structure, adolescents are 51.21% more likely to have high or medium behavioural risk, holding all other covariates constant.
- For every unit increase in their emotional risk score, adolescents are 0.74% more likely to have high or medium behavioural risk, holding all other covariates constant.

- For every unit increase in their self-esteem score, adolescents are 4.76% less likely to have high or medium behavioural risk, holding all other covariates constant.

The researchers therefore concluded that based on the model, a profile for an adolescent with high behavioural risk is a male who intends to drop out of school, comes from a single-parent home, possibly has high emotional risk scores and low self esteem scores.

The researchers did well by putting forth a profile; However, they could incorporate a level of significance and a confidence interval for their adjusted O.R.s and interpretations.

Model Diagnostics

The researchers conducted diagnostics in two key ways: performing cross-validation and using a variety of tools to conduct model and coefficient diagnostics. They performed cross-validation by splitting the sample randomly into 10 sub-samples. The model was then fit on the overall dataset and each of the sub samples. For various aspects of model diagnostics, they were careful to use more than one test statistic and explain the limitations of each test where possible.

Significance of the entire model

The entire model was found to be statistically significant in both the sub-samples and overall sample.

The Likelihood ratio, the Score test and the Wald test all had p-values that were less than 0.0001

Because the Chi-Square Test of Proportional Odds assumption had a p-value of 0.6548, there was no need to fit a second model.

Furthermore, the Hosmer and Lemeshow test was conducted on the individual models that were estimating p_1 and $p_1 or p_2$ and both were found statistically significant - the p-values of both test statistics were > 0.4.

Significance of individual covariate coefficients

All covariate coefficients were found to be statistically significant except EMOTION - this had a p-value that was greater than 0.05 (0.5211).

However, FAMILY and DROPOUT covariate coefficients were found to be statistically significant in 9 out of the 10 sub-samples.

Validation of predictive probabilities

A number of measures of association were used such as Kendall's tau-a, Goodman-Kruskal's Gamma, Sommers' D statistic and the c statistic. Goodman-Kruskal's Gamma was 0.548, indicating that fewer errors were made in predicting the probabilities using the model than by using random chance. The c statistic value was 0.769, indicating that the model correctly categorized 76.9% of the possible pairs of adolescents; therefore, the model is better than random chance.

Handling of missing cases

The researchers discovered that EMOTION was the variable that had most of the missing cases and posited that this may have contributed to the insignificance of its individual coefficient. Missing data was imputed using SPSS using the EM method, a statistical software, and the model was fitted on the modified data. However, the statistical significance of individual coefficients was the same as the model that was fir on data that had the observations with missing data completely omitted.

The researchers conducted thorough diagnostics, going so far as to select 36 cases, comparing their estimated probabilities to the actual category of the dependent variable that each observation was classified in. This was good because it enabled them to further defend the model and evaluate individual covariate coefficient results.

Reference

Peng, C.Y.J. and Nichols, R.N., 2003. Using multinomial logistic models to predict adolescent behavioral risk. *Journal of Modern Applied Statistical Methods*, 2(1), p.16.

Problem 8

In at most four pages, briefly discuss generalized estimating equations (GEE) and generalized linear mixed models (GLMM) using an example in each case.

Solution for Question 8

Generalized Estimating Equations (GEE)

The same subject in a study may change over time; therefore, there may be correlation between the successive measurements and this makes the assumption that the observations in Y are independent become implausable.

Furthermore, a subject may have a more similar measurement to another subject within the same group compared to a subject from a different group

The two scenarios above make the observations become correlated with one another. To handle this kind of data, the correlation structure within the data can be modelled. This forms **generalized estimating equations**.

The following example will be used to illustrate how these equations are applied.

Example: Per Capita Expenditure

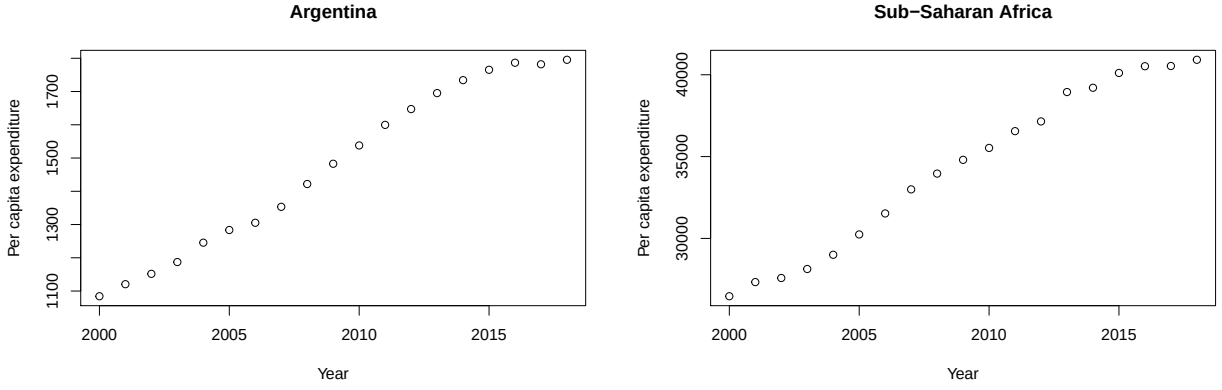
Data source: https://bit.ly/3blw0be

Per Capita Expenditure (PCE) is the total market value of all purchases in a country divided by that country's total population. It is a measure used to assess the disposable income that the average citizen has in a given country/territory.

Data on per capita expenditure for 120 countries and territories was collected by the World Bank for the period 2000-2018. These countries were grouped into 7 regions:

- East Asia and Pacific
- Europe and Central Asia
- Latin America and the Caribbean
- Middle East and North Africa
- Other High Income
- South Asia
- Sub-Saharan Africa

When we take an example of plotting Argentina's and Sub-Saharan Africa's graph below, we see that the individual observations are correlated over time at an observational level and at a group level:



If there are N number of subjects(in this case, territories) with n_i measurements for subject i,

let $\boldsymbol{y}_i$ denote the vector of responses for each territory and $\boldsymbol{y}$ denote the vector of responses for all territories.

The length of $\boldsymbol{y}$ will be $\sum_{i=i}^{N} n_i$

In our example, n_i measurements for each of the territories is equal $\therefore n_i = n_j = 19$

$$E(\boldsymbol{y}_i) = \boldsymbol{\mu}_i$$

$$g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i^T \boldsymbol{\beta}$$

$$\boldsymbol{D} = \frac{\partial \boldsymbol{\mu}_i}{\partial \beta_j}$$

Therefore, the **generalized estimating equations** are:

i.

$$\boldsymbol{U} = \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0$$

where

ii.

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i \boldsymbol{A}_i^{\frac{1}{2}} \phi$$

where

$\boldsymbol{A}$ is a matrix with elements, $var(\boldsymbol{y}_{ik})$ $\boldsymbol{R}_i$ is a correlation matrix for $\boldsymbol{y}_i$ and ϕ is a constant for overdispersion

The GEEs are solved iteratively as follows:

1. Start with $\boldsymbol{R}_i = \boldsymbol{I}$ and $\phi = 1$
2. Estimate $\boldsymbol{\beta}$ by solving equations in *i.*
3. Calculate fitted values $\widehat{\boldsymbol{\mu}}_i = g^{-1}(\boldsymbol{X}_i^T)\widehat{\boldsymbol{\beta}}$
4. Get the residuals $\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i$
5. Use the resulting residuals to estimate $\boldsymbol{A}_i$, $\boldsymbol{R}_i$ and ϕ
6. Repeat steps 2, 3, 4 and 5 until convergence is achieved.

If the data is continous, correlation is used but when it is binary, odds ratios can be used. Furthermore, for longitudinal data, a sandwich estimator must be used for $var(\boldsymbol{\beta})$ given as

$\boldsymbol{V}_s(\widehat{\boldsymbol{\beta}}) = \Im^{-1} \boldsymbol{C} \Im^{-1}$

where

$\Im = \sum_{i=1}^{N} \boldsymbol{D}_i^T \widehat{\boldsymbol{V}}_i^{-1} \boldsymbol{D}_i$ and $\boldsymbol{C} = \sum_{i=1}^{N} \boldsymbol{D}_i^T \widehat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)^T \boldsymbol{D}_i$

Application of this approach on our data (with the Wald test statistic being applied) gives the following statistically significant coefficients:

term	estimate	std.error	statistic	p.value
(Intercept)	1240.473	299.9409	17.104213	0.0000354
regionEurope and Central Asia	2498.871	628.8525	15.790287	0.0000708
regionLatin America and the Caribbean	2051.187	487.8455	17.678520	0.0000262
regionMiddle East and North Africa	1246.261	614.8555	4.108394	0.0426709
regionOther High Income	17595.925	1283.1750	188.040949	0.0000000
year2016	1312.906	663.2658	3.918247	0.0477647
year2017	1432.561	692.4333	4.280265	0.0385573
year2018	1586.839	732.4177	4.694052	0.0302672

Upon applying the Wald test to the fitted model, it was found to be statistically significant $\because \alpha_{cal} << 0.05$ as shown below:

Df	X2	P(>|Chi|)
132	5597.768	0

Generalized Linear Mixed Models (GLMM)

An Ordinary Least Squares Regression Model is of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$$
$$E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X\beta}$$

When we want to use generalized linear models, we expand the class of the response variable from strictly normal distribution to any distribution within the exponential family of distributions, which includes the normal distribution. This is achieved by applying a transformation to the expected value of $\boldsymbol{y}$, $\boldsymbol{\mu}$ as shown:

$$g(\boldsymbol{\mu}) = \boldsymbol{X\beta}$$

Therefore, a generalized linear mixed models is where the expected value of the response variable of a generalized linear model has both fixed and random effects i.e.

$$g(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$$

where $\boldsymbol{\beta}$ are the fixed effects and the random effects of the model ($\boldsymbol{\varepsilon}$) are split into two, random effects between subjects, $\boldsymbol{Zu}$ and random effects within subjects, $\boldsymbol{e}$.

To make the calculations more easier to handle, particular pairs of distributions are used called *conjugate distributions*. This approach is therefore similar to Bayesian analysis.

Examples of conjugate distributions include:

- Normal distribution for $\boldsymbol{y}|\boldsymbol{u}$ and Normal distribution for $\boldsymbol{u}$
- Poisson distribution for $\boldsymbol{y}|\boldsymbol{u}$ and Gamma distribution for $\boldsymbol{u}$
- Binomial distribution for $\boldsymbol{y}|\boldsymbol{u}$ and Beta distribution for $\boldsymbol{u}$
- Binomial distribution for $\boldsymbol{y}|\boldsymbol{u}$ and Normal distribution for $\boldsymbol{u}$

Example: Alexa Reviews

Data source: https://go.aws/2WkOfZp

When Amazon customers purchase an Amazon Alexa, a voice-controlled virtual assistant, some of them may choose to leave a review and give the product a rating from one to 5. The data below consists of data surrounding 3,150 customers that left a review. It contains the following variables:

- *rating*: The actual rating that the customer gave a product on a scale of 1 to 5
- *variation*: The outer colour/pattern that the product had
- *date*: The time the customer posted the review
- *love_dummy*: The review contained one of the 10 most frequent verbs, love.
- *great_dummy*: The review contained one of the 10 most frequent verbs, great.
- *like_dummy*: The review contained one of the 10 most frequent verbs, like.
- *easy_dummy*: The review contained one of the 10 most frequent verbs, easy.
- *works_dummy*: The review contained one of the 10 most frequent verbs, works.
- *good_dummy*: The review contained one of the 10 most frequent verbs, good.
- *doesnt_dummy*: The review contained one of the 10 most frequent verbs, doesn't.
- *quality_dummy*: The review contained one of the 10 most frequent verbs, quality.
- *better_dummy*: The review contained one of the 10 most frequent verbs, better.
- *well_dummy*: The review contained one of the 10 most frequent verbs, well.

(Code for all the dummy variables: Yes=1, No=0)

A generalized linear mixed model was fitted onto the data by maximum likelihood (Laplace Approximation) with *date* and *variation* as the random effects. This resulted in the following statistically significant fixed-effect coefficients (based on the z test statistic):

effect	group	term	estimate	std.error	statistic	p.value
fixed	NA	(Intercept)	2.1610479	0.2323574	9.300533	0.0000000
fixed	NA	love_dummy1	2.3389317	0.3959834	5.906641	0.0000000
fixed	NA	great_dummy1	1.3529132	0.3168865	4.269393	0.0000196
fixed	NA	like_dummy1	0.8365031	0.3036561	2.754771	0.0058733
fixed	NA	easy_dummy1	2.8309525	1.0160859	2.786135	0.0053341
fixed	NA	good_dummy1	0.8456607	0.4088853	2.068210	0.0386202
fixed	NA	doesnt_dummy1	-1.4875350	0.3624097	-4.104567	0.0000405

Upon applying the Likelihood Ratio test to the fitted model, it was found to be statistically significant $\because \alpha_{cal} << 0.05$ as shown below:

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Question8Model2_null	3	1237.590	1255.756	-615.7952	1231.590	NA	NA	NA
Question8Model2	15	1110.755	1201.582	-540.3774	1080.755	150.8356	12	0

References

1. Dobson, A.J. and Barnett, A.G., 2018. *An introduction to generalized linear models.* CRC press.
2. Gbur, E.E., Stroup, W.W., McCarter, K.S., Durham, S., Young, L.J., Christman, M., West, M. and Kramer, M., 2020. *Analysis of generalized linear mixed models in the agricultural and natural resources sciences* (Vol. 156). John Wiley & Sons.