The University of Hong Kong
School of Public Health

CMED6100/MMPH6002/CMED7100
Introduction to Biostatistics (Semester I)
Practical 4 Suggested solution

Part 1. Linear regression

1. Fit the following two linear regression models to examine the estimated effects of household income on self-reported health. Complete Table 1.

   Model 1: explain the self-reported health score with household income group only (as a categorical predictor).

   In R Commander, categories of a categorical variable are arranged in alphabetical or numerical order by default. In order to reorder the categories and set the correct category as reference group, we need to recode the following variables: household income, gender and smoking status. The category coded with the smallest number will be the reference group. We don't need to recode chronic disease variable as the "no doctor-diagnosed chronic conditions reported" category is recognized as reference group by default.

   [Data → Manage variables in active dataset → Recode variables→ Variables to recode: hinc; New variable name: hinc.rc; Enter recode directives:
   "<$5,000" = 1
   "$5,000-$9,999" = 2
   "$10,000-$19,999" = 3
   "$20,000-$39,999" = 4
   "$40,000 or above" = 5]
   [Statistics → Fit models → Linear Model → score ~ hinc.rc]
   [Models → Confidence intervals]

   Model 2: explain the self-reported health score with household income group (as a categorical predictor), age, gender, smoking status and chronic disease.

   [Data → Manage variables in active dataset → Recode variables→ Variables to recode: smoker; New variable name: smoker.rc; Enter recode directives:
   "never" = 1
   "former" = 2
   "current smoker" = 3]
   [Data → Manage variables in active dataset → Recode variables→ Variables to recode: male; New variable name: male.rc; Enter recode directives:
   "male" = 1
   "female" = 2]
   [Statistics → Fit models → Linear Model → score ~ hinc.rc + age + male.rc + smoker.rc + chronic]
   [Models → Confidence intervals]

Table 1. Linear regression model for the self-reported health status

	Model 1		Model 2	
	β	95% CI	β	95% CI
Monthly household income				
<$5,000	0		0	
$5,000 – $9,999	3.241	(-0.253, 6.736)	1.012	(-2.537, 4.560)
$10,000 – $19,999	3.173	(0.100, 6.246)	0.375	(-2.850, 3.600)
$20,000 – $39,999	5.142	(2.046, 8.237)	2.281	(-1.002, 5.564)
$40,000 or above	5.791	(2.284, 9.298)	3.001	(-0.648, 6.650)
Age (per year)			-0.023	(-0.079, 0.033)
Male			0	
Female			-2.296	(-4.015, -0.577)
Smoking Status				
Never smoker			0	
Former smoker			-4.153	(-8.043, -0.263)
Current smoker			-1.227	(-3.377, 0.923)
Presence of doctor-diagnosed chronic conditions				
No			0	
Yes			-3.636	(-5.916, -1.355)
Adjusted R^2	0.026		0.078	

2. We could assess the goodness of fit of Models 1 and 2 using adjusted R^2. Which model has a better fit? Why?

   Model 2 has a better fit to the data because the adjusted R^2 for Model 2 (0.078) is larger than the adjusted R^2 for Model 1 (0.026).

3. Is the relation between household income and self-reported health score statistically significant in Model 2? Why?

   There were no significant association between household income and self-reported health score at 5% level of significance, after adjusting for other variables (the 95% CIs all include 0).

4. Based on Model 2, what is the predicted difference in self-reported health score for a male versus a female in the same household income group, age, smoking status and chronic disease?

   The predicted self-reported health score for a male is 2.296 points higher than a female with the same household income group, age, chronic disease and smoking status.

Part 2. Logistic regression

5. We could assess the association of self-reported health score with other covariates by fitting the logistic regression model. Create a new variable “gdhealth” to classify the people reported with poor to fair, defined as a health score of at most 20 (set “gdhealth” = 0) versus good to excellent, defined as a health score higher than 20 (set “gdhealth” = 1).

   Fit the following logistic regression models to examine the estimated effects of household income on self-reported health. Complete Table 2.

   Model 3: explain the dichotomized self-reported health status (variable gdhealth) with household income group only (as a categorical predictor).

   [Data → Manage variables in active dataset → Recode variables→ Variables to recode: score; New variable name: gdhealth; Enter recode directives:
   0:20 = 0
   else = 1]

   [Statistics → Fit models → Generalized Linear Model → gdhealth ~ hinc.rc; Enter name for model: GLM.3; Family: binomial; Link function: logit]

   To obtain the 95% confidence intervals for exponentiated coefficient estimates, we can use “confint.default” function in R by typing the command on R Console panel:
   exp(confint.default(GLM.3))
   Note: GLM.3 is the model name which is indicated in the ‘Model’ section on the right of the “View data set” tab in the R Commander window.

   Model 4: explain the dichotomized self-reported health status (variable gdhealth) with household income group (as a categorical predictor), age, gender, smoking status and chronic disease.

   [Statistics → Fit models → Generalized Linear Model → gdhealth ~ hinc.rc + age + male.rc + smoker.rc + chronic; Enter name for model: GLM.4; Family: binomial; Link function: logit]

   Again, to obtain the 95% confidence intervals, we type the command on R Console panel:
   exp(confint.default(GLM.4))

Table 2. Logistic regression models for the self-reported health status

	Model 3		Model 4	
	OR	95% CI for OR	OR	95% CI for OR
Monthly household income				
<$5,000	1.00		1.00	
$5,000 – $9,999	2.173	(0.919, 5.137)	1.488	(0.597, 3.707)
$10,000 – $19,999	1.516	(0.727, 3.162)	0.923	(0.408, 2.089)
$20,000 – $39,999	3.180	(1.472, 6.872)	1.896	(0.810, 4.439)
$40,000 or above	3.789	(1.505, 9.538)	2.353	(0.871, 6.353)
Age (per year)			0.992	(0.977, 1.007)
Male			1.00	
Female			0.690	(0.429, 1.108)
Smoking Status				
Never smoker			1.00	
Former smoker			0.944	(0.333, 2.681)
Current smoker			0.847	(0.473, 1.514)
Presence of doctor-diagnosed chronic conditions				
No			1.00	
Yes			0.517	(0.286, 0.933)
AIC	489.73		488.03	

6. We could assess the goodness of fit of Models 3 and 4 using AIC. Which model has a better fit? Why?

   The AIC for Model 4 (488.03) is lower than the AIC for Model 3 (489.73). However, a difference in AIC of less than 2 suggests that the two models are essentially indistinguishable in goodness of fit. The difference in AIC between Model 3 and Model 4 is smaller than 2, so the two models are essentially not much different in terms of model fit.

7. Is the relation between chronic condition and dichotomized self-reported health status statistically significant in Model 4? Why?

   The presence of chronic condition is significantly associated with health status at 5% level of significance after adjusting for other variables, 95% CI = (0.286, 0.933) does not include 1.

8. Dichotomizing the outcome variable ‘self-reported health score’ may result in loss of information. Which factors are significant in Model 2 but not in Model 4?

   Former smoker and sex are significant factors after adjusting for other variables in Model 2 but not in Model 4.

Part 3. Categorizing continuous variables

9. We could access the age effect on health status by categorizing age into subgroups. Create a new variable “agegp” with values from 1 to 4 representing ‘15-30’, ‘31-45’, ‘46-60’, ‘60+’ respectively.

   Fit the following logistic regression model to examine the estimated effects of household income on self-reported health. Complete Table 3.

   Model 5: explain the dichotomized self-reported health status (variable gdhealth) with household income group (as a categorical predictor), age (as a categorical predictor), gender, smoking status and chronic disease.

   [Data → Manage variables in active dataset → Recode variables→ Variables to recode: age; New variable name: agegp; Enter recode directives:
   15:30 = 1
   31:45 = 2
   46:60 = 3
   else = 4]

   [Statistics → Fit models → Generalized Linear Model → gdhealth ~ hinc.rc + agegp + male.rc + smoker.rc + chronic; Enter name for model: GLM.5; Family: binomial; Link function: logit]

   To obtain the 95% confidence intervals, we type the command on R Console panel:
   exp(confint.default(GLM.5))

Table 3. Logistic regression models for the self-reported health status

	Model 5	
	OR	95% CI for OR
Monthly household income		
<$5,000	1.00	
$5,000 – $9,999	1.529	(0.608, 3.843)
$10,000 – $19,999	0.954	(0.416, 2.188)
$20,000 – $39,999	1.907	(0.808, 4.501)
$40,000 or above	2.435	(0.889, 6.670)
Age 15-30	1.00	
Age 31-45	0.723	(0.393, 1.331)
Age 46-60	0.535	(0.274, 1.048)
Age 60+	0.641	(0.287, 1.428)
Male	1.00	
Female	0.670	(0.415, 1.081)
Smoking Status		
Never smoker	1.00	
Former smoker	0.917	(0.323, 2.603)
Current smoker	0.848	(0.472, 1.522)
Presence of doctor-diagnosed chronic conditions		
No	1.00	
Yes	0.517	(0.288, 0.929)

10. Compared with young adults (aged 15-30), does old adults (aged 60+) have a lower or higher odds of reporting better health? Why?

    Compared with younger adults, older adults have a lower odds of reported better health, but the difference is insignificant (OR = 0.641, 95% CI = 0.287 to 1.428).