School of Mathematics and Statistics

Te Kura Mātai Tatauranga

STAT 292 **Assignment 4: Due Thursday, 1 June 2023 at 11:59 PM**

Note: Your assignment can be typed or handwritten (and scanned). Be sure to submit your assignment as a PDF and follow the instructions specified on the course Nuku page. Where calculations are performed in `R`, you must include relevant code and output with your answer to receive credit.

Assignments that are submitted late will receive a mark of 0 unless illness, bereavement or other substantial causes occur and have been discussed with the course coordinator.

1. (25 marks)

   The table below presents data from the Framingham Heart Study, which explores risk factors for cardiovascular disease. It is of interest to understand whether systolic blood pressure (SBP), measured in millimetres of mercury (mmHg), is associated with incidence of hypertension.

SBP Range (mmHg)	SBP Midpoint (mmHg)	Hypertensive	Not Hypertensive
< 120	100	15	1264
120 - < 130	125	81	866
130 - < 140	135	160	570
140 - < 180	160	896	218
≥ 180	200	165	5

   a. Fit the logistic regression model

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X,$$

   where X denotes systolic blood pressure as represented by stated midpoints for presented ranges and $p(X)$ denotes the probability of hypertension. Attach `R` code used to fit the logistic regression model, and also include summary output for the model. (4 marks)

   b. Carry out an appropriate goodness-of-fit test to determine whether the model provides a good fit to the data. State the hypotheses, and give the test statistic and the p-value of the test. What do you conclude at the $\alpha = 0.05$ significance level? (5 marks)

   c. Give estimates of β_0 and β_1 (to at least 4dp). (2 marks)

   d. Interpret the association between systolic blood pressure (as measured numerically by midpoints of systolic blood pressure ranges) and incidence of hypertension using the odds ratio (to at least 3dp). Demonstrate how the odds

ratio is calculated from summary output in part (a). Additionally, provide a 95% confidence interval for the odds ratio (to at least 3dp). (5 marks)

e. Find the predicted probability (to at least 4dp) of hypertension for a person with a systolic blood pressure of 125 mmHg. (3 marks)

f. Find the fitted count of incidence of hypertension (to at least 2dp) for people with a systolic blood pressure of 125 mmHg. Also find the fitted count of those without hypertension (to at least 2dp) for people with a systolic blood pressure of 125 mmHg. (3 marks)

g. Test

$$\begin{aligned} \mathcal{H}_0 : \beta_1 &= 0 \\ \mathcal{H}_1 : \beta_1 &\neq 0 \end{aligned}$$

using the Wald statistic. Give the test statistic and the p-value of the test. What do you conclude at the $\alpha = 0.05$ significance level? (3 marks)

2. (25 marks)

Now we consider the same data as for Question 1 but treating systolic blood pressure range as categorical. Additional information was collected for each person on whether they were a smoker and whether they were diabetic. Incidence of hypertension based on smoker status, diabetes status, and systolic blood pressure range are as presented in the table below.

Smoker (W)	Diabetic (X)	SBP (Y)	Hypertensive Yes	Hypertensive No
Yes	Yes	< 120	1	1
		$120 - < 130$	0	0
		$130 - < 140$	0	0
		$140 - < 180$	5	0
		≥ 180	0	0
	No	< 120	9	715
		$120 - < 130$	33	459
		$130 - < 140$	79	274
		$140 - < 180$	367	94
		≥ 180	55	3
No	Yes	< 120	0	1
		$120 - < 130$	0	2
		$130 - < 140$	1	1
		$140 - < 180$	10	1
		≥ 180	2	0
	No	< 120	5	547
		$120 - < 130$	48	405
		$130 - < 140$	80	295
		$140 - < 180$	514	123
		≥ 180	108	2

a. Fit the logit model

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) \;=\; \beta_0 + \beta_i^W + \beta_j^X + \beta_k^Y + \beta_{ij}^{WY},$$

where p_{ijk} is the probability of hypertension when the smoker status (W) is at level i, diabetes status (X) is at level j, and systolic blood pressure (Y) is at level k. Attach `R` code used to fit the logit model, and also include summary output for the model. (4 marks)

b. Interpret any interaction effects represented in this model. What do these interaction effects mean or assume? (Note that you are not being asked to interpret coefficients. You are strictly being asked to interpret what it means for specific variables to interact in the context of this problem.) (4 marks)

c. Is the model fit in part (a) a saturated model? Why or why not? (Note that rows in the table with no observations [i.e., 0 for both hypertensive and not hypertensive columns] are not counted towards the number of logits being estimated in the model.) (2 marks)

d. Carry out a goodness-of-fit test for the model presented in part (a). Give the test statistic and the p-value of the test. What do you conclude at the $\alpha = 0.05$ significance level? (3 marks)

e. Carry out a model comparison of the model fit in part (a) with the model

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) \;=\; \beta_0 + \beta_j^X + \beta_k^Y,$$

(Be sure to present relevant `R` code and output.) Write down the hypotheses to be tested, test statistic, distribution of the test statistic under the null hypothesis, p-value, and conclusion at the $\alpha = 0.05$ significance level. Be sure to provide a qualitative explanation of what this result tells us about the importance of smoker status. (7 marks)

f. For the model presented in part (e), compare the odds of hypertension for diabetics with the odds of hypertension for non-diabetics using an odds ratio, and provide a precise interpretation of this odds ratio. Give a 95% confidence interval for the odds ratio. (5 marks)