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

1

a.

Code:

Untitled1*

Source on Save | Run | Source

```r
# Store the data in vectors for the variables.
midpoint.sbp <- c(100, 125, 135, 160, 200)
range.sbp <- c("< 120", "120 - < 130","130 - < 140", "140 - < 180", ">= 180")
hypertensive <- c(15, 81, 160, 896, 165)
not.hypertensive <- c(1264, 866, 570, 218, 5)

# Fit the logistic regression model.
logistic.regression <- glm(cbind(hypertensive, not.hypertensive) ~ midpoint.sbp, family =
                           "binomial")
summary(logistic.regression)
```

10:29 (Top Level) R Script

Console | Terminal | Background Jobs

R 4.2.2 · ~/

summary output:

```
Call:
glm(formula = cbind(hypertensive, not.hypertensive) ~ midpoint.sbp,
    family = "binomial")

Deviance Residuals:
      1       2       3       4       5
 1.3823  -1.0218  -0.4817   1.0448  -3.2739

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.066143   0.438132  -34.39   <2e-16 ***
midpoint.sbp   0.102508   0.003032   33.81   <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: 2623.394  on 4  degrees of freedom
Residual deviance:   14.997  on 3  degrees of freedom
AIC: 46.81

Number of Fisher Scoring iterations: 4

>
```

β_0

β_1

b.

Hypotheses:

null hypothesis: Model M provides a good fit to the data

alternative hypothesis: Model M does not provide a good fit the data

where the model M denotes the logistic regression model that we fit.

Under the null hypothesis, the deviance is given by:

$$G^2 \approx 14.997$$

which has approximately a χ^2_3 distribution.

p-value is given by:

$$\text{p-value} \approx P(\chi^2_3 > 14.997) \approx 0.0018$$

```
11
12  p.value <- pchisq(14.997, df = 3, lower.tail = FALSE)
13  p.value
```

13:8 (Top Level) R Script

Console | Terminal | Background Jobs

R 4.2.2 · ~/

```
Deviance Residuals:
      1       2       3       4       5
 1.3823  -1.0218  -0.4817   1.0448  -3.2739

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.066143   0.438132  -34.39   <2e-16 ***
midpoint.sbp  0.102508   0.003032   33.81   <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: 2623.394  on 4  degrees of freedom
Residual deviance:   14.997  on 3  degrees of freedom
AIC: 46.81

Number of Fisher Scoring iterations: 4

>
> p.value <- pchisq(14.997, df = 3, lower.tail = FALSE)
> p.value
[1] 0.001819214
```

Wald = z value = deviance = goodness of fit test statistic:

$$z^* = \frac{\beta_1}{SE(\hat{\beta}_1)} = \frac{0.102508}{0.003032} \approx 33.81$$

As the p-value is 0.001819 (4 sf), which is less 0.05, have significant evidence to reject H_0 & conclude that the logistic regression model does not provide a good fit to this sample data.

A4_q1.R*

```
1  # Store the data in vectors for the variables.
2  midpoint.sbp <- c(100, 125, 135, 160, 200)
3  range.sbp <- c("< 120", "120 - < 130","130 - < 140", "140 - < 180", ">= 180")
4  hypertensive <- c(15, 81, 160, 896, 165)
5  not.hypertensive <- c(1264, 866, 570, 218, 5)
6
7  # Fit the logistic regression model.
8  logistic.regression <- glm(cbind(hypertensive, not.hypertensive) ~ midpoint.sbp, family =
9                               "binomial")
10 summary(logistic.regression)
11
12 p.value <- pchisq(14.997, df = 3, lower.tail = FALSE)
13 p.value
14 # deviance and residual degrees of freedom
15 G.2 <- logistic.regression$deviance
16 residual.df <- logistic.regression$df.residual
17 # Goodness-of-fit p-value.
18 p_value <- pchisq(q = G.2, df = residual.df, lower.tail = FALSE)
19 p_value
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2623.394  on 4  degrees of freedom
Residual deviance:   14.997  on 3  degrees of freedom
AIC: 46.81

Number of Fisher Scoring iterations: 4

> p.value <- pchisq(14.997, df = 3, lower.tail = FALSE)
> p.value
[1] 0.001819214
> # deviance and residual degrees of freedom
> G.2 <- logistic.regression$deviance
> residual.df <- logistic.regression$df.residual
> # Goodness-of-fit p-value.
> p_value <- pchisq(q = G.2, df = residual.df, lower.tail = FALSE)
> p_value
[1] 0.001819136
```

same p-value

deviance goodness of-fit

c. From summary output, we observe that:

$\hat{\beta}_0 \approx -15.066143$ (6 dp)

$\hat{\beta}_1 \approx 0.102508$ (6 dp)

(label in the summary output (a))

d. The association between systolic blood pressure and incidence of hypertension can be interpreted through $\exp(\beta_1)$, which is $\exp(0.102508) = 1.107946165444$ or 1.108 (3dp). In particular, we estimate that an increase in the midpoint of systolic blood pressure is associated with a multiplicative change of 1.108 (1.101, 1.115) (3 dp) in the odds of incidence of hypertension. 95% CI

A4_q1.R*

Source on Save | Run | Source

Go back to the previous source location (⌘F9)

```
2  midpoint.sbp <- c(100, 125, 135, 160, 200)
3  range.sbp <- c("< 120", "120 - < 130","130 - < 140", "140 - < 180", ">= 180")
4  hypertensive <- c(15, 81, 160, 896, 165)
5  not.hypertensive <- c(1264, 866, 570, 218, 5)
6
7  # Fit the logistic regression model.
8  logistic.regression <- glm(cbind(hypertensive, not.hypertensive) ~ midpoint.sbp, family =
9                               "binomial")
10 summary(logistic.regression)
11
12 p.value <- pchisq(14.997, df = 3, lower.tail = FALSE)
13 p.value
14
15 # Produce estimate for odds ratios.
16 exp(logistic.regression$coefficients)
17
18 # Produce 95% confidence intervals corresponding to odds ratios.
19 exp(confint.default(logistic.regression))
```

14:1 (Top Level) R Script

Console | Terminal | Background Jobs

R 4.2.2 · ~/

```
    Null deviance: 2623.394  on 4  degrees of freedom
Residual deviance:   14.997  on 3  degrees of freedom
AIC: 46.81

Number of Fisher Scoring iterations: 4

>
> p.value <- pchisq(14.997, df = 3, lower.tail = FALSE)
> p.value
[1] 0.001819214
>
> # Produce estimate for odds ratios.
> exp(logistic.regression$coefficients)
 (Intercept) midpoint.sbp
2.863237e-07 1.107946e+00
>
> # Produce 95% confidence intervals corresponding to odds ratios.
> exp(confint.default(logistic.regression))
                    2.5 %       97.5 %
(Intercept)  1.213161e-07 6.757657e-07
midpoint.sbp 1.101382e+00 1.114549e+00
>
```

e.

$$\hat{p}(x) \approx \frac{\exp(-15.066143 + 0.102508x)}{1+\exp(-15.066143 + 0.102508x)}$$

$$\hat{p}(125) \approx \frac{\exp(-15.066143 + 0.102508 \times 125)}{1+\exp(-15.066143 + 0.102508 \times 125)}$$

$$\approx 0.0951 \ (4 \text{ dp})$$

A4_q1.R*

Source on Save | Run | Source

```
4  hypertensive <- c(15, 81, 160, 896, 165)
5  not.hypertensive <- c(1264, 866, 570, 218, 5)
6
7  # Fit the logistic regression model.
8  logistic.regression <- glm(cbind(hypertensive, not.hypertensive) ~ midpoint.sbp, family =
9                               "binomial")
10 summary(logistic.regression)
11
12 p.value <- pchisq(14.997, df = 3, lower.tail = FALSE)
13 p.value
14
15 # Produce estimate for odds ratios.
16 exp(logistic.regression$coefficients)
17
18 # Produce 95% confidence intervals corresponding to odds ratios.
19 exp(confint.default(logistic.regression))
20
21 # Produce predicted probabilities of a person of 125 mmHg systolic blood pressure with hypertension
22 predict(logistic.regression, newdata = data.frame(midpoint.sbp = 125), type = "response")
```

20:1 (Top Level) R Script

Console | Terminal | Background Jobs

R 4.2.2 · ~/

```
> p.value <- pchisq(14.997, df = 3, lower.tail = FALSE)
> p.value
[1] 0.001819214
>
> # Produce estimate for odds ratios.
> exp(logistic.regression$coefficients)
 (Intercept) midpoint.sbp
2.863237e-07 1.107946e+00
>
> # Produce 95% confidence intervals corresponding to odds ratios.
> exp(confint.default(logistic.regression))
                   2.5 %       97.5 %
(Intercept)  1.213161e-07 6.757657e-07
midpoint.sbp 1.101382e+00 1.114549e+00
>
> # Produce predicted probabilities of a heart attack for someone smoking 25 cigarettes per day.
> predict(logistic.regression, newdata = data.frame(midpoint.sbp = 125), type = "response")
         1
0.09512352
>
```

f.

from e. $= \hat{p}(125) \approx \dfrac{\exp(-15.066143 + 0.102508 \times 125)}{1 + \exp(-15.066143 + 0.102508 \times 125)}$

$\approx$ 0.0951 (4 dp)

with hypertension:

↳ fitted count for presence of hypertension for X = 125 mmHg of systolic blood pressure:

- fitted count $\approx (81 + 866) \times 0.0951$

$\approx$ 90.0597 people

Thus 947 people with a systolic blood pressure of X = 125 mmHg, we would expect around 90.0597 people to have incidence of hypertensive

without hypertension:

↳ fitted count for absence of hypertension for X = 125 mmHg. of systolic blood pressure:

- fitted count $\approx (81+866) - 90.0597 \approx 856.9403$ people

g. Wald Test :

$$z^* = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.102508}{0.003032} \approx 33.80870712$$

$$\approx 33.81 \text{ (2 dp)}$$

p-value :

$$\text{p-value} = 2 \times P(Z > |z^*|)$$

$$\approx 2 \times P(Z > 33.81)$$

$$\approx 2 \times \underbrace{2.076 \times 10^{-248}}_{0}$$

$$\approx 2 \times 0 \approx 0$$

p-value is $< 2 \times 10^{-16}$ (in summary output), which is almost 0. Since p-value is less than 0.05 significant level, we reject H_0, meaning that there is a significant association between systolic blood pressure & the incidence of hypertension. The systolic blood pressure is a significant predictor of hypertension in this data.

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)

(2) a.

```r
data <- data.frame(
  smoker = c("Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes",
        "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"),
  diabetic = c("Yes", "Yes", "Yes", "Yes", "Yes", "No", "No", "No", "No", "No",
        "Yes", "Yes", "Yes", "Yes", "Yes", "No", "No", "No", "No", "No"),
  sbp = c("<120", "120 - <130", "130 - <140", "140 - <180", ">=180",
        "<120", "120 - <130", "130 - <140", "140 - <180", ">=180",
        "<120", "120 - <130", "130 - <140", "140 - <180", ">=180"),
  Hypertensive = c(1, 0, 0, 5, 0, 9, 33, 79, 367, 55, 0, 0, 1, 10, 2, 5, 48, 80, 514, 108),
  Not.Hypertensive = c(1, 0, 0, 0, 0, 715, 459, 274, 94, 3, 1, 2, 1, 1, 0, 547, 405, 295, 123, 2)
)
print(data)
model <- glm(cbind(Hypertensive, Not.Hypertensive) ~ factor(smoker) * factor(sbp) + factor(diabetic), data = data, family = binomial)
summary(model)
```

```
Call:
glm(formula = cbind(Hypertensive, Not.Hypertensive) ~ factor(smoker) *
    factor(sbp) + factor(diabetic), family = binomial, data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.26894  -0.04937   0.00000   0.02284   1.76710

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -4.7025     0.4493 -10.466  < 2e-16 ***
factor(smoker)Yes                       0.4229     0.5507   0.768   0.4425
factor(sbp)>=180                        8.6959     0.8432  10.313  < 2e-16 ***
factor(sbp)120 - <130                   2.5543     0.4746   5.382 7.35e-08 ***
factor(sbp)130 - <140                   3.3965     0.4665   7.281 3.32e-13 ***
factor(sbp)140 - <180                   6.1284     0.4603  13.315  < 2e-16 ***
factor(diabetic)Yes                     1.4463     0.7253   1.994   0.0462 *
factor(smoker)Yes:factor(sbp)>=180      -1.5076     1.0789  -1.397   0.1623
factor(smoker)Yes:factor(sbp)120 - <130 -0.9072     0.5993  -1.514   0.1301
factor(smoker)Yes:factor(sbp)130 - <140 -0.3605     0.5792  -0.622   0.5336
factor(smoker)Yes:factor(sbp)140 - <180 -0.4829     0.5715  -0.845   0.3981
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2640.5804  on 16  degrees of freedom
Residual deviance:    5.7691  on  6  degrees of freedom
AIC: 80.896

Number of Fisher Scoring iterations: 5
```

b.

```
factor(smoker)Yes:factor(sbp)>=180       -1.5076     1.0789  -1.397   0.1623
factor(smoker)Yes:factor(sbp)120 - <130  -0.9072     0.5993  -1.514   0.1301
factor(smoker)Yes:factor(sbp)130 - <140  -0.3605     0.5792  -0.622   0.5336
factor(smoker)Yes:factor(sbp)140 - <180  -0.4829     0.5715  -0.845   0.3981
---
```

There are no interaction effects that are statistically significant at the significance level of 0.05, as can be seen from the summary output of the all p-values associated with the interaction effects between smoking status (Yes) and various levels of systolic blood pressure (sbp) are greater than 0.05, which implies that there is no significant evidence of interaction between smoker status (Yes) and the different levels of systolic blood pressure. As a result of this context, the effect of smoker status on hypertension does not appear to vary significantly depending on the specific ranges of systolic blood pressure. Although these interaction effects are not statistically significant, it does not mean that there is no relationship or interaction between the variables.

c.

It's not a saturated model because 1) residual deviance is 6 df, which is not 0 df (0 df means that it is a saturated model and 2) non-reductant parameter:

$$1 + 1 + 4 + 1 + (1\times 4) = 11$$

(5 systolic blood level - 1)

(2 level of diab. - 1)

(2 level of smoking - 1)

number of the observation is not equal to 11

↳ so not saturated model

d. test stat: $G^2(M) \approx 5.7691460$

The residual df is 6, so $G^2(M) \sim X^2_6$.

p-value:

$$\text{p-value} \approx P(X^2_6 > 5.7691460)$$

$$\approx 0.4495 \text{ (4 sf)}$$

```
print(data)
model <- glm(cbind(Hypertensive, Not.Hypertensive) ~ factor(smoker) * factor(sbp) + factor(diabetic), data = data, family = binomial)
summary(model)

p.value <- pchisq(5.7691, df = 6, lower.tail = FALSE)
p.value

G.2 <- model$deviance
residual.df <- model$df.residual
# Goodness-of-fit p-value.
pvalue <- pchisq(q = G.2, df = residual.df, lower.tail = FALSE)
c(G.2, residual.df, pvalue)
```

11:89 (Top Level)

Console | Terminal | Background Jobs

R 4.2.2 · ~/

```
factor(sbp)130 - <140                    3.3965     0.4665   7.281 3.32e-13 ***
factor(sbp)140 - <180                    6.1284     0.4603  13.315  < 2e-16 ***
factor(diabetic)Yes                      1.4463     0.7253   1.994   0.0462 *
factor(smoker)Yes:factor(sbp)>=180       -1.5076     1.0789  -1.397   0.1623
factor(smoker)Yes:factor(sbp)120 - <130  -0.9072     0.5993  -1.514   0.1301
factor(smoker)Yes:factor(sbp)130 - <140  -0.3605     0.5792  -0.622   0.5336
factor(smoker)Yes:factor(sbp)140 - <180  -0.4829     0.5715  -0.845   0.3981
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2640.5804  on 16  degrees of freedom
Residual deviance:    5.7691  on  6  degrees of freedom
AIC: 80.896

Number of Fisher Scoring iterations: 5
```

```
>
> p.value <- pchisq(5.7691, df = 6, lower.tail = FALSE)
> p.value
[1] 0.4495465
```

↳ equal p-value : 0.4495 (to check)

```
>
> G.2 <- model$deviance
> residual.df <- model$df.residual
> # Goodness-of-fit p-value.
> pvalue <- pchisq(q = G.2, df = residual.df, lower.tail = FALSE)
> c(G.2, residual.df, pvalue)
[1] 5.7691460 6.0000000 0.4495411
>
```

G.2 → test statistic; 6 → df; p-value

p-value = 0.4495

- p-value is more than 0.05 significant level, which implies that there is no significant evidence to reject null hypothesis at 0.05 significant level & conclude that the model does fit the data.

e. reduced model (M_1)

full model (M_2)

Test hypotheses :

H_0: Additional terms in model M_2 can be deleted

H_1: Additional terms in model M_2 can't be deleted

A4_Q2.R*

```
  smoker = c("Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes",
        "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"),
  diabetic = c("Yes", "Yes", "Yes", "Yes", "Yes", "No", "No", "No", "No", "No",
        "Yes", "Yes", "Yes", "Yes", "Yes", "No", "No", "No", "No", "No"),
  sbp = c("<120", "120 - <130", "130 - <140", "140 - <180", ">=180",
        "<120", "120 - <130", "130 - <140", "140 - <180", ">=180",
        "<120", "120 - <130", "130 - <140", "140 - <180", ">=180",
        "<120", "120 - <130", "130 - <140", "140 - <180", ">=180"),
  Hypertensive = c(1, 0, 0, 5, 0, 9, 33, 79, 367, 55, 0, 0, 1, 10, 2, 5, 48, 80, 514, 108),
  Not.Hypertensive = c(1, 0, 0, 0, 0, 715, 459, 274, 94, 3, 1, 2, 1, 1, 0, 547, 405, 295, 123, 2)
)
print(data)
model <- glm(cbind(Hypertensive, Not.Hypertensive) ~ factor(smoker) * factor(sbp) + factor(diabetic), data = data, family = "binom
summary(model)

M1.model <- glm(cbind(Hypertensive, Not.Hypertensive) ~ factor(diabetic) + factor(sbp), data = data, family = "binomial")
M2.model <- glm(cbind(Hypertensive, Not.Hypertensive) ~ factor(smoker) * factor(sbp) + factor(diabetic),  data = data, family = "b
# Carry out a model comparison
anova(M1.model, M2.model, test = "Chisq")

20:42 (Top Level) R Script
```

Console | Terminal | Background Jobs

R 4.2.2 · ~/Desktop/

```
    Null deviance: 2640.5804  on 16  degrees of freedom
Residual deviance:    5.7691  on  6  degrees of freedom
AIC: 80.896

Number of Fisher Scoring iterations: 5

>
> M1.model <- glm(cbind(Hypertensive, Not.Hypertensive) ~ factor(diabetic) + factor(sbp), data = data, family = "binomial")
> M2.model <- glm(cbind(Hypertensive, Not.Hypertensive) ~ factor(smoker) * factor(sbp) + factor(diabetic),  data = data, family = "binom
ial")
> # Carry out a model comparison
> anova(M1.model, M2.model, test = "Chisq")
Analysis of Deviance Table

Model 1: cbind(Hypertensive, Not.Hypertensive) ~ factor(diabetic) + factor(sbp)
Model 2: cbind(Hypertensive, Not.Hypertensive) ~ factor(smoker) * factor(sbp) +
    factor(diabetic)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        11    12.3306
2         6     5.7691  5   6.5614   0.2554
>
```

- Test statistic:

$$G^2 = 12.3306 - 5.7691$$

$$\approx 6.56 \ (3 \text{ sf})$$

- Under H_0, $G^2 \sim \chi^2_5$ ← $df = 11 - 6 = 5$

- p-value:

$$\text{p-value} \approx P(\chi^2_5 > 6.5614)$$

$$\approx 0.2554$$

- As the p-value exceeds 0.05, we can conclude that we have insufficient evidence to reject H_0.
↳ The additional terms in Model M_2 can be deleted, leading to the form of the reduced model that excludes the smoker. Which implies that the smoker does not significant impacted on the outcome.

f. As you can see in the R output below, the odds ratio for hypertension in diabetics are associated with multiplicative change of approximately 4.476 (1.058319, 18.92717) in the odds of those non-diabetics.

A4_Q2.R*

Source on Save | Run | Source

```
8          "<120", "120 - <130", "130 - <140", "140 - <180", ">=180",
9          "<120", "120 - <130", "130 - <140", "140 - <180", ">=180"),
10   Hypertensive = c(1, 0, 0, 5, 0, 9, 33, 79, 367, 55, 0, 0, 1, 10, 2, 5, 48, 80, 514, 108),
11   Not.Hypertensive = c(1, 0, 0, 0, 0, 715, 459, 274, 94, 3, 1, 2, 1, 1, 0, 547, 405, 295, 123, 2)
12 )
13 print(data)
14 model <- glm(cbind(Hypertensive, Not.Hypertensive) ~ factor(smoker) * factor(sbp) + factor(diabetic), data = data, family = "binom
15 summary(model)
16
17 M1.model <- glm(cbind(Hypertensive, Not.Hypertensive) ~ factor(diabetic) + factor(sbp), data = data, family = "binomial")
18 M2.model <- glm(cbind(Hypertensive, Not.Hypertensive) ~ factor(smoker) * factor(sbp) + factor(diabetic),  data = data, family = "b
19 # Carry out a model comparison
20 anova(M1.model, M2.model, test = "Chisq")
21
22 # Produce estimate for odds ratios.
23 exp(M1.model$coefficients)
24
25 # Produce 95% confidence intervals corresponding to odds ratios.
26 exp(confint.default(M1.model))
```

26:31 (Top Level) R Script

Console | Terminal | Background Jobs

R 4.2.2 · ~/Desktop/

```
   factor(diabetic)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        11    12.3306
2         6     5.7691  5   6.5614   0.2554
>
> # Produce estimate for odds ratios.
> exp(M1.model$coefficients)
         (Intercept)  factor(diabetic)Yes     factor(sbp)>=180 factor(sbp)120 - <130 factor(sbp)130 - <140
        1.177482e-02         4.475598e+00         2.776389e+03          7.898750e+00          2.371062e+01
factor(sbp)140 - <180
        3.444700e+02
>
> # Produce 95% confidence intervals corresponding to odds ratios.
> exp(confint.default(M1.model))
                              2.5 %       97.5 %
(Intercept)           7.075541e-03 1.959516e-02
factor(diabetic)Yes   1.058319e+00 1.892717e+01
factor(sbp)>=180      9.959504e+02 7.739679e+03
factor(sbp)120 - <130 4.521659e+00 1.379809e+01
factor(sbp)130 - <140 1.383674e+01 4.063046e+01
factor(sbp)140 - <180 2.026959e+02 5.854069e+02
>
```

⇒ 1.058319 – 18.92717 95% CI