STAT292
Assignment 4
Isabelle Southon
300597453

Question 1

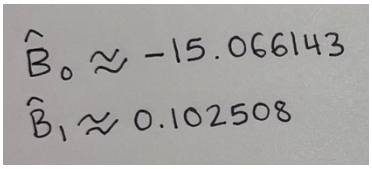
a) Fit the logistic regression model

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

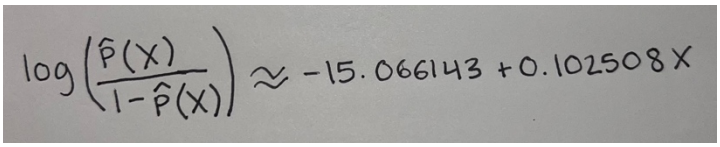
```
### Question 1
SBP.midpoint <- c(100, 125, 135, 160, 200)
hypertensive <- c(15, 81, 160, 896, 165)
not.hypertensive <- c(1264, 866, 570, 218, 5)

hypertensive.model <- glm(cbind(hypertensive, not.hypertensive) ~ SBP.midpoint,
                          family = "binomial")
summary(hypertensive.model)
```

```
Call:
glm(formula = cbind(hypertensive, not.hypertensive) ~ SBP.midpoint,
    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 ***
SBP.midpoint   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
```

- Estimated regression coefficients



- Fitted logistic regression equation



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?

H_0 : Model M provides a good fit to the data
H_1 : Model M does not provide a good fit to the data

Under the null hypothesis, the deviance is given by:

$$G^2(M) \approx 14.997$$

which has approximately a X^2_3 distribution. The p-value is given by:

$$\text{p-value} \approx P(X^2_3 > 14.997) \approx 0.00182$$

As the p-value is less than a = 0.05, we reject H_0, and we conclude that the logistic regression model does not provide a good fit to the cardiovascular disease data. The test statistics is 14.997 and there are 3 degrees of freedom.

```
# Deviance Goodness-of-fit test

# Deviance and residual degrees of freedom
G <- hypertensive.model$deviance
residual.df <- hypertensive.model$df.residual

# Goodness-of-fit p-value
p.value <- pchisq(q = G, df = residual.df,
                  lower.tail = FALSE)

c(G, residual.df , p.value)

# Double check the p-value
p.value <- pchisq(14.997, df = 3, lower.tail = FALSE)
p.value
```

```
> c(G, residual.df , p.value)
[1] 14.997091496  3.000000000  0.001819136
> # Double check the p-value
> p.value <- pchisq(14.997, df = 3, lower.tail = FALSE)
> p.value
[1] 0.001819214
```

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

$$\hat{\beta}_0 \approx -15.0661$$
$$\hat{\beta}_1 \approx 0.1025$$

d) Interpret the association between systolic blood pressure and incidence of hypertension using the odds ratio Demonstrate how the odds ratio is calculated from summary output in part (a). Additionally, provide a 95% confidence interval for the odds ratio

```
# Produce estimate for odds ratios.
round(exp(hypertensive.model$coefficients),3)

# Produce 95% confidence intervals corresponding to odds ratios.
round(exp(confint.default(hypertensive.model)),3)
```

```
  (Intercept) SBP.midpoint
        0.000        1.108
```

95% confidence interval

```
                2.5 % 97.5 %
(Intercept)     0.000  0.000
SBP.midpoint    1.101  1.115
```

The association between systolic blood pressure (mmHg) and incidence of hypertension can be interpreted through exp(β1). In particular, we estimate that an increase in the systolic blood pressure by 1 is associated with multiplicative change of 1.1079 (1.101, 1.115) in the odds of the incidence of hypertension. (i.e., each systolic blood pressure level increase is estimated to increase the odds of hypertension by 11%)

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

$$\hat{P}(125) \approx \frac{\exp(-15.066143 + 0.102508 \times 125)}{1 + \exp(-15.066143 + 0.102508 \times 125)} \approx 0.0951$$

```
# 1e) Predicted probability
predict(hypertensive.model, newdata = data.frame(SBP.midpoint = 125), type = "response")
```

```
          1
0.09512352
```

The predicted probability of having hypertension when a person's systolic blood pressure is 125 mmHg is approximately 0.0951.

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.

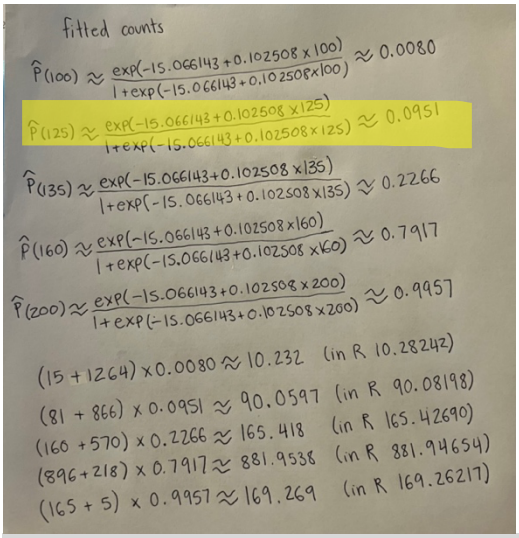
Find the fitted count of incidence of hypertension for people with a systolic blood pressure of X = 125 mmHg.
Fitted count $\approx (81 + 866) \times 0.0951 \approx 90.0597\ people$

Thus, for 947 people with a systolic blood pressure of 125 mmHg, we would expect approximately 90.06 to have hypertension based on this model.

The fitted count of those without hypertension for people with a systolic blood pressure of X = 125 mmHg follows,

Fitted count $\approx (81 + 866) - 90.0597 = 856.9403$ *people*

Thus, for 947 people with a systolic blood pressure of 125 mmHg, we would expect approximately 90.06 to be hypertensive, and 856.94 to be not hypertensive based on this model.




```
# 1f) Fitted counts
n <- hypertensive + not.hypertensive
fitted.counts <- n * predict(hypertensive.model, type = "response")
cbind(hypertensive, fitted.counts)
```

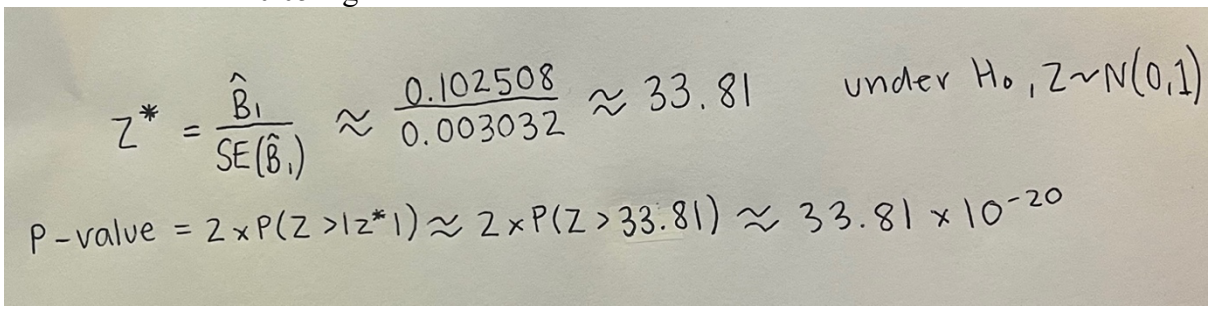
	hypertensive	fitted.counts
1	15	10.28242
2	81	90.08198
3	160	165.42690
4	896	881.94654
5	165	169.26217

g) Test:

H_0: $B_1 = 0$

H_1: $B_1 \neq 0$

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?



The p-value is 33.81×10^{-20}, which is well below $\alpha = 0.05$ (or any reasonable significance level). Therefore we reject H_0 at the 5% significance level because the p-value is essentially zero. Therefore, there is strong evidence that suggests that incidence of hypertension is associated with systolic blood pressure (SBP).

Question 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} denotes the probability of hypertension when:
W (smoker status) is at level i
X (diabetes status) is at level j
Y (systolic blood pressure) is at level k

```
# 2a) Input the logit model data
SBP <- rep(c("<120", "120-<130", "130-<140", "140-<180", ">=180"), times = 4)
length(SBP)
hypertension.yes <- c(1, 0, 0, 5, 0, 9, 33, 79, 367, 55, 0, 0, 1, 10, 2, 5, 48,
                      80, 514, 108) # yes hypertension
length(hypertension.yes)
hypertension.no <- c(1, 0, 0, 0, 0, 715, 459, 274, 94, 3, 1, 2, 1, 1, 0, 547,
                     405, 295, 123, 2) # no hypertension
length(hypertension.no)
diabetes <- rep(c("Yes", "No"), each = 5, times = 2)
length(diabetes)
smoker <- rep(c("Yes","No"), each = 10, times = 1)
length(smoker)

# Fit the logit model
hypertension.model <- glm(cbind(hypertension.yes, hypertension.no) ~ factor(SBP) *
                            factor(smoker) + factor(diabetes), family = "binomial")
summary(hypertension.model)
```

```
Call:
glm(formula = cbind(hypertension.yes, hypertension.no) ~ factor(SBP) *
    factor(smoker) + factor(diabetes), family = "binomial")

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(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(smoker)Yes                    0.4229     0.5507   0.768   0.4425
factor(diabetes)Yes                  1.4463     0.7253   1.994   0.0462 *
factor(SBP)>=180:factor(smoker)Yes   -1.5076     1.0789  -1.397   0.1623
factor(SBP)120-<130:factor(smoker)Yes -0.9072     0.5993  -1.514   0.1301
factor(SBP)130-<140:factor(smoker)Yes -0.3605     0.5792  -0.622   0.5336
factor(SBP)140-<180:factor(smoker)Yes -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) Interpret any interaction effects represented in this model

H_0: there is no interaction present
H_1: there is interaction present
The reference level is measured at each level of systolic blood pressure (mmHg), and the person being a non-smoker. To interpret the interaction effects, we measure the interaction at each systolic blood pressure level by comparing it to the reference level.

At the systolic blood pressure level 180 and above, there is no significant interaction with yes smoker when compared to the reference level. This is because there is a p-value of 0.1623, which is greater than any reasonable significance level.

At the systolic blood pressure levels between 120 and less than 130, there is no significant interaction with yes smoker when compared to the reference level. This is because there is a p-value of 0.1301, which is greater than any reasonable significance level.

At the systolic blood pressure levels between 130 and less than 140, there is no significant interaction with yes smoker when compared to the reference level. This is because there is a p-value of 0.5336, which is greater than any reasonable significance level.

At the systolic blood pressure levels between 140 and less than 180, there is no significant interaction with yes smoker when compared to the reference level. This is because there is a p-value of 0.3981, which is greater than any reasonable significance level.

According to the model, there is no significant interaction between any of the systolic blood pressure levels and with (yes) smoking. This means that for all levels of systolic blood pressure, we fail to reject H_0 and conclude that there is no interaction present.

c) Is the model fit in part (a) a saturated model? Why or why not?

No, it is not a saturated model because it does not include all possible interactions. In the model we have an interaction between smoker status (W) and systolic blood pressure (Y), but there are also three additional interactions that could be tested. These are the interaction between:

- smoker status (W) and diabetes (X),
- diabetes (X) and systolic blood pressure (Y)
- smoker status (W), diabetes (X) and systolic blood pressure (Y)

A model that is saturated is said to have 0 residual degrees of freedom, and this model is not saturated because it has a residual deviance of 5.7691 on 6 degrees of freedom.

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?

H_0: Model M provides a good fit to the data
H_1: Model M does not provide a good fit to the data
Under the null hypothesis, the deviance is given by
$G^2(M) \approx 5.769$
Which approximately has a X^2_6 distribution. The p-value is given by
p-value $\approx P(X^2_6 > 5.7691) \approx 0.449$

The test statistic is 5.7691, there are 6 degrees of freedom and the p-value of 0.449 is greater than the 5% significance level, therefore we fail to reject H_0, and conclude that the logit model provides a good fit to the data.

```
# 2d) goodness of fit
G2 <- hypertension.model$deviance
residual.df2 <- hypertension.model$df.residual
p.value <- pchisq(q = G2, df = residual.df2, lower.tail = FALSE)
c(G2, residual.df2, p.value)
```

```
[1] 5.7691460 6.0000000 0.4495411
```

e) Carry out a model comparison

```
# 2e)
model_1 <- glm(cbind(hypertension.yes, hypertension.no) ~ factor(diabetes) +
                 factor(SBP), family = "binomial")

model_2 <- glm(cbind(hypertension.yes, hypertension.no) ~ factor(SBP) *
                            factor(smoker) + factor(diabetes), family = "binomial")

anova(model_1, model_2, test = "Chisq")
```

```
Model 1: cbind(hypertension.yes, hypertension.no) ~ factor(diabetes) +
    factor(SBP)
Model 2: cbind(hypertension.yes, hypertension.no) ~ factor(SBP) * factor(smoker) +
    factor(diabetes)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        11    12.3306
2         6     5.7691  5   6.5614   0.2554
```

Hypotheses:
H_0: the additional terms in Model M2 can be deleted
H_1: the additional terms in Model M2 cannot be deleted.

Test statistic:
$G^2 = G^2(M_1) - G^2(M_2)$
$\approx 6.5614 - 5.7691 = 0.7923$,

Which follows an approximate X^2_5 distribution under H_0. The p-value is given by,
p-value $\approx P(X^2_5 > 0.7923) \approx 0.2554$

As the p-value exceeds a significance level of $\alpha = 0.05$, we have insufficient evidence to reject the null hypothesis, meaning that the additional terms in Model M2 can be deleted, leading to the form of the reduced model that excludes smoker status as an explanatory variable.

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)

STAT292
Assignment 4
Isabelle Southon
300597453

We estimate that the odds of incidence of hypertension for those who have diabetes is 4.476 (1.058, 18.927) times that of those who do not have diabetes. Since the confidence interval is above 1, this is a significantly higher odds of incidence of hypertension for those who have diabetes than for those who do not have diabetes. The 95% confidence interval for the odds ratio is (1.058, 18.927). This model also accounts for the systolic blood pressure level ranges.

```
# Produce estimate for odds ratios.
round(exp(model_1$coefficients),3)

# Produce 95% confidence intervals corresponding to odds ratios.
round(exp(confint.default(model_1)),3)
```

```
> round(exp(model_1$coefficients),3)
       (Intercept) factor(diabetes)Yes    factor(SBP)>=180 factor(SBP)120-<130 factor(SBP)130-<140 factor(SBP)140-<180
             0.012                4.476            2776.389               7.899              23.711             344.470
> round(exp(confint.default(model_1)),3)
                      2.5 %   97.5 %
(Intercept)           0.007    0.020
factor(diabetes)Yes   1.058   18.927
factor(SBP)>=180    995.950 7739.679
factor(SBP)120-<130   4.522   13.798
factor(SBP)130-<140  13.837   40.630
factor(SBP)140-<180 202.696  585.407
```