Homework 4

2023-10-12

Problem 1

The Poisson distribution may be useful to model events such as

a) Time interval between bus arrivals
b) The price of the plane ticket
c) The time of the first goal during the soccer
d) The number of bus arriving to a bus station between 12:00 and 16:00

We can use a Poisson distribution to predict or explain the number of events occurring within a given interval of time or space. That is why d) is correct.

Problem 2

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(glm2)

data_1 <- read.csv("diabetes.csv")

# the Logistic regression model
model_1 <- glm(Outcome ~ Insulin + BMI + Pregnancies, data = data_1, family = binomial)

summary(model_1)
```

```
##
## Call:
## glm(formula = Outcome ~ Insulin + BMI + Pregnancies, family = binomial,
##     data = data_1)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.9721  -0.8797  -0.5904   1.0927   2.5256
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.5245216  0.4627101  -9.778  < 2e-16 ***
## Insulin      0.0016953  0.0007011   2.418   0.0156 *
## BMI          0.0940904  0.0128626   7.315 2.57e-13 ***
## Pregnancies  0.1721690  0.0279865   6.152 7.66e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 957.38  on 744  degrees of freedom
## Residual deviance: 845.31  on 741  degrees of freedom
## AIC: 853.31
##
## Number of Fisher Scoring iterations: 4
```

The intercept represents the log-odds of having diabetes when all predictor variables (Insulin, BMI, and Pregnancies) are equal to 0.

The coefficient for "Insulin" represents how a one-unit increase in Insulin affects the log-odds of having diabetes, holding other variables constant. In this case, it's positive (0.0016953), indicating that as Insulin increases, the log-odds of having diabetes also increase. This is a significant effect as indicated by the p-value (0.0156) and Signif. Code *.

The coefficient for "BMI" (0.0940904) represents how a one-unit increase in BMI affects the log-odds of having diabetes while holding other variables constant. It's also positive, suggesting that higher BMI is associated with an increased likelihood of having diabetes. This effect is highly significant with a very small p-value (2.57e-13) and Signif. Code ***.

The coefficient for "Pregnancies" (0.1721690) represents how a one-unit increase in the number of pregnancies affects the log-odds of having diabetes while controlling for other variables. It's also positive, indicating that more pregnancies are associated with a higher likelihood of having diabetes. This effect is significant with a p-value of (7.66e-10) and Signif. Code ***.

So, all three predictor variables have p-values significantly less than 0.05, which indicates their statistical significance in explaining the likelihood of diabetes.

Find the minimum and maximum values of predictions and discuss how well the model predicts the probability of having diabetes.

```
predicted_1 <- predict(model_1, newdata = data_1, type = "response")
data_1$Prediction = predicted_1
#View(data_1)
min(predicted_1)
```

```
## [1] 0.01072366
```

```
max(predicted_1)
```

```
## [1] 0.9080677
```

We predicted that the patient can have a diabetes with the minimum probability of 0.01072366 and the actual outcome is 0 which means that patient does not have diabetes. Now, same for the maximal probability of 0.9080677, we see that the actual outcome is 1 which means that patient does have diabetes. So, it is clear that the model predicts the probability of having diabetes very well.

What is the probability for person with median BMI and Insulin with 0 pregnancy to have diabetes?

```
# type = "response" argument to get the predicted probability of having diabetes.
predict(model_1, newdata = data.frame(Insulin = median(data_1$Insulin), BMI = median(data_1$BMI), Pregna
```

```
##         1
## 0.1896021
```

What is odd ratio for patients with maximum BMI and maximum Insulin with 4 pregnancies?

```
# type = "link" to get the predicted log-odds.
log_odds <- predict(model_1, newdata = data.frame(Insulin = max(data_1$Insulin), BMI = max(data_1$BMI),

# We calculate the odds ratio by exponentiating the log-odds.
exp(log_odds)
```

```
##        1
## 49.99096
```

The odds ratio represents the multiplicative change in the odds of the event (having diabetes) associated with a one-unit increase in the predictor variable.

Problem 3

```
data_2 <- read.csv("Armenian_pub.csv")
# the Poisson regression model
data_2$Freq <- factor(data_2$Freq, levels = 0:2, labels = c("Rare","Several times a month", "Several ti
poisson_model <- glm(Age ~ Income + WTS + Freq, data = data_2, family = poisson)

summary(poisson_model)
```

```
##
## Call:
## glm(formula = Age ~ Income + WTS + Freq, family = poisson, data = data_2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.45451  -0.20909  -0.00567   0.23178   0.49324
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.933e+00  3.466e-02  84.604   <2e-16 ***
## Income                      8.550e-08  1.591e-07   0.537    0.591
## WTS                         1.131e-06  4.594e-06   0.246    0.805
## FreqSeveral times a month -7.812e-04  4.075e-02  -0.019    0.985
## FreqSeveral times a week   6.612e-03  5.858e-02   0.113    0.910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 10.430  on 153  degrees of freedom
## Residual deviance: 10.004  on 149  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 757.61
##
## Number of Fisher Scoring iterations: 3
```

The estimated coefficient for the "Intercept" is 2.933e+00. This represents the expected log count of "Age" when all other predictors are zero (Income = 0, WTS = 0, and FreqSeveral times a month = "Rare", FreqSeveral times a week = "Rare"). The very small p-value (2e-16) indicates that this intercept is highly significant and Signif. Code ***.

The estimated coefficient for "Income" is approximately 8.550e-08. However, the p-value for Income is 0.591, indicating that it is not statistically significant at conventional significance levels (e.g., 0.05). This suggests that changes in income are not associated with significant changes in the count of Age.

The estimated coefficient for "WTS" is approximately 1.131e-06, which is also close to zero. The p-value is 0.805, indicating that WTS is not statistically significant in predicting age. Similar to income, this predictor does not appear to be statistically significant.

The coefficient for "Frequency (Several times a month)" is approximately -7.812e-04, and the p-value is 0.985. This suggests that the frequency of visiting the pub "Several times a month" is not statistically significant in predicting age.

The coefficient for "Frequency (Several times a week)" is approximately 6.612e-03, and the p-value is 0.910. Like the previous frequency category, visiting the pub "Several times a week" is not statistically significant in predicting age.

In summary, the intercept is highly significant, suggesting that there is a significant relationship between the intercept and the variable Age. However, the other predictor variables (Income, WTS, and both levels of Frequency) are not statistically significant in explaining the variation in "Age." These variables do not appear to be strongly associated with the Age based on this model.

Obtain predictions from the Poisson regression model. Find the minimum and maximum values and discuss how well the model predicts age.

```
predicted_2 <- predict(poisson_model, newdata = data_2, type = "response", na.action = "na.exclude")
predicted_2
```

```
##        1        2        3        4        5        6        7        8
## 18.96683 18.90068 19.00314 18.82664 19.20341 19.09125 18.78409 18.80535
##        9       10       11       12       13       14       15       16
## 19.12149 19.06965 18.97067 18.92257 18.94490 19.19473 18.99929 18.91774
##       17       18       19       20       21       22       23       24
## 18.80535 19.02940 18.88403 19.14465 18.92680 19.11713 19.03041 18.98830
##       25       26       27       28       29       30       31       32
## 18.90068 18.86978 18.96683 19.01364 18.89235 19.30598 18.79472 18.86268
##       33       34       35       36       37       38       39       40
## 18.88403 19.01364 19.01364 18.93342 18.90304 19.13399 18.89400 18.82005
##       41       42       43       44       45       46       47       48
## 19.11285 18.96634 19.14465 18.91774 18.98166 18.84135 18.92680 18.93967
##       49       50       51       52       53       54       55       56
## 18.86839 19.35955 18.92780 18.93303 19.01504 18.93966 18.87499 18.79877
```

```
##       57       58       59       60       61       62       63       64
## 18.95969 18.92895 18.99739 18.94112 18.88167 18.85772 18.92680 18.85797
##       65       66       67       68       69       70       71       72
## 18.91202 19.10139 18.85201 18.82664 18.84795 18.86268 19.12778 19.33717
##       73       74       75       76       77       78       79       80
## 19.19473 18.92372 19.21605 18.86928 18.77752 19.09652 18.88403 19.27697
##       81       83       84       85       86       87       88       89
## 19.12921 18.89660 19.11863 19.22126 18.98115 18.94349 18.88467 18.86432
##       90       91       92       93       94       95       96       97
## 19.27166 19.21596 18.98115 20.51026 18.88592 18.93966 19.46732 18.89496
##       98       99      100      101      102      103      104      105
## 19.04618 19.05620 18.87360 18.79472 20.51026 18.98115 18.90068 19.02990
##      106      107      108      109      111      112      113      114
## 18.86928 18.85365 19.12778 18.92207 18.90207 18.82664 18.92348 19.75624
##      115      116      117      118      120      121      122      123
## 19.18700 18.82005 18.82005 19.22995 19.15421 18.94872 18.91685 18.96493
##      124      125      126      127      128      129      130      131
## 18.93253 19.05863 19.32392 19.06774 19.20535 18.88403 19.06965 18.91774
##      132      133      134      135      136      137      138      139
## 19.07878 18.83209 18.80535 19.15421 18.81624 19.31531 18.91774 18.90590
##      141      142      143      144      145      146      147      148
## 18.82664 19.30904 18.98442 18.96493 19.04806 19.09886 18.84135 18.91774
##      149      150      151      152      153      154      155      156
## 18.93826 18.79472 18.87360 19.04132 19.00981 18.95446 18.83615 18.76285
##      157      158
## 19.01364 19.02006
```

```r
min(predicted_2)
```

```
## [1] 18.76285
```

```r
max(predicted_2)
```

```
## [1] 20.51026
```

Here, we got our predicted minimum age of 18.76285, but in the data we have a lot of ages being 17. Also, for the maximum age our predicted is 20.51026, but again we have values for age being equal to 21. From here, we see that our model does not predict age well.

Find the expected age of visitor who has 200.000 income, 3000 WTS and is coming to Pub Several times a week.

```r
# Set the predictor values
income_value <- 200000
wts_value <- 3000
# Frequency = "Several times a week"
freq_value <- 2

# Calculate the expected age
expected_age_c <- exp(coef(poisson_model)["(Intercept)"] + (income_value * coef(poisson_model)["Income"]
expected_age_c
```

```
## (Intercept)
##    19.42146
```

Find the expected age of visitor who has 100.000 income, 5000 WTS and is coming to Pub rare.

```
# Set the predictor values
income_value <- 100000
wts_value <- 5000

# Calculate the expected age
expected_age_d <- exp(coef(poisson_model)["(Intercept)"] + (income_value * coef(poisson_model)["Income"]
expected_age_d
```

```
## (Intercept)
##    19.04618
```

Calculate the probability that at least one of the visitors from c) and d) will be older than 25. (Considering they are independent from each other). P(A B) = P(A) + P(B) - P(A B)

```
# Calculate the probabilities of being older than 25 for each visitor
probability_visitor_c <- 1 - ppois(25, lambda = expected_age_c, lower.tail = TRUE)
probability_visitor_d <- 1 - ppois(25, lambda = expected_age_d, lower.tail = TRUE)

# Calculate the probability of the intersection (both not being older than 25)
probability_intersection <- probability_visitor_c * probability_visitor_d

# Calculate the probability that at least one of them is older than 25 using the inclusion-exclusion pr
probability_at_least_one_older <- probability_visitor_c + probability_visitor_d - probability_intersecti

# Print the result
probability_at_least_one_older
```

```
## [1] 0.156348
```