HW8

1. Separation (12 points)

a. (4 points)

Assume y ~ Binomial(n,p), what is the MLE of p? Using your MLE what is the point estimate and confidence interval for p when y = n?

b. (4 points)

Now consider a logistic regression model

$$\begin{aligned} y_i &\sim Bernoulli(p_i) \\ logit(p_i) &= \beta_0 + \beta_1 I(zipcode = group2) \end{aligned}$$

If all of the y_i values for group 1 are equal to 0 and all the y_i values for group 2 are equal to 1, what should be the MLE for β_0? Hint: you can use the `invlogit` function in the `arm` package to explore this.

c. (4 points)

Diagnose the model

```
seattle %>% group_by(zipcode) %>% summarise(mean(waterfront))
```

```
## # A tibble: 2 x 2
##   zipcode `mean(waterfront)`
##   <fct>                <dbl>
## 1 98070                    1
## 2 98102                    0
```

```
glm_separation <- glm(waterfront ~ zipcode, data = seattle, family = binomial)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
display(glm_separation)
```

```
## glm(formula = waterfront ~ zipcode, family = binomial, data = seattle)
##              coef.est coef.se
## (Intercept)     26.57 72693.36
## zipcode98102   -53.13 80574.03
## ---
##   n = 129, k = 2
##   residual deviance = 0.0, null deviance = 124.0 (difference = 124.0)
```

2. Diagnostics / Simulation (8 points)

a. (4 points)

Simulate data and fit a logistic regression model to this data. Comment on your results.

b. (4 points)

Repeat this simulation many times and create diagnostics plots to look at the residuals. Present one plot and comment on the results

3. Titanic Prediction (10 points)

The `titanic` package in R contains survival data on passengers from the Titanic. Use the titanic_train dataset to fit a logistic regression model for passenger survival. Then report your classification error on the titanic_test dataset. Look at a few different model specifications and summarize what you found to work best.

```
library(titanic)
tibble(titanic_train)
```

```
## # A tibble: 891 x 1
##    titanic_train$P... $Survived $Pclass $Name $Sex   $Age $SibSp $Parch $Ticket
##                <int>     <int>   <int> <chr> <chr> <dbl>  <int>  <int> <chr>
##  1                  1         0       3 Brau... male     22      1      0 A/5 21...
##  2                  2         1       1 Cumi... fema...    38      1      0 PC 175...
##  3                  3         1       3 Heik... fema...    26      0      0 STON/O...
##  4                  4         1       1 Futr... fema...    35      1      0 113803
##  5                  5         0       3 Alle... male     35      0      0 373450
##  6                  6         0       3 Mora... male     NA      0      0 330877
##  7                  7         0       1 McCa... male     54      0      0 17463
##  8                  8         0       3 Pals... male      2      3      1 349909
##  9                  9         1       3 John... fema...    27      0      2 347742
## 10                 10         1       2 Nass... fema...    14      1      0 237736
## # ... with 881 more rows, and 3 more variables: $Fare <dbl>, $Cabin <chr>,
## #   $Embarked <chr>
```

```
tibble(titanic_test)
```

```
## # A tibble: 418 x 1
##    titanic_test$Pa... $Pclass $Name $Sex   $Age $SibSp $Parch $Ticket $Fare $Cabin
##                <int>   <int> <chr> <chr> <dbl>  <int>  <int> <chr>   <dbl> <chr>
##  1                892       3 Kell... male   34.5      0      0 330911   7.83 ""
##  2                893       3 Wilk... fema...  47        1      0 363272   7    ""
##  3                894       2 Myle... male   62        0      0 240276   9.69 ""
##  4                895       3 Wirz... male   27        0      0 315154   8.66 ""
##  5                896       3 Hirv... fema...  22        1      1 3101298 12.3  ""
##  6                897       3 Sven... male   14        0      0 7538     9.22 ""
##  7                898       3 Conn... fema...  30        0      0 330972   7.63 ""
##  8                899       2 Cald... male   26        1      1 248738  29    ""
##  9                900       3 Abra... fema...  18        0      0 2657     7.23 ""
## 10                901       3 Davi... male   21        2      0 A/4 48... 24.2  ""
## # ... with 408 more rows, and 1 more variable: $Embarked <chr>
```