Quiz 3

Neev Shaw

October 25, 2025

Question 1

[1]:
```
Donner = read.csv("DonnerParty.csv")
attach(Donner)
fix(Donner)
summary(Donner)
```

```
    Name                Age           Survive           sex
 Length:88          Min.   : 1.00   Min.   :0.0000   Length:88
 Class :character   1st Qu.: 7.00   1st Qu.:0.0000   Class :character
 Mode  :character   Median :16.50   Median :1.0000   Mode  :character
                    Mean   :20.19   Mean   :0.5568
                    3rd Qu.:29.25   3rd Qu.:1.0000
                    Max.   :70.00   Max.   :1.0000
     Sex            Status              Hired
 Min.   :0.0000   Length:88          Min.   :0.0000
 1st Qu.:0.0000   Class :character   1st Qu.:0.0000
 Median :1.0000   Mode  :character   Median :0.0000
 Mean   :0.6023                      Mean   :0.1591
 3rd Qu.:1.0000                      3rd Qu.:0.0000
 Max.   :1.0000                      Max.   :1.0000
```

As introduced in the beginning of the quiz this data refers to the Donner Party - a group of emigrants who faced harsh conditions during the winter of 1846-1847 in the Sierra Nevada Mountains. The columns in this dataset correspond to the following:

- Name: Name of the emigrant
- Age: Age of the emigrant
- Survive: binary variable corresponding to whether or not the emigrant survived the winter (survived = 1)
- sex (lowercase): binary text variable corresponding to sex of the emigrant
- Sex (uppercase): binary variable corresponding to sex (but Male = 1)
- Status: a varaible with 3 classifications correspodning to status of emigrant (Family, Single, Hired)
- Hired: binary variable where Hired = 1 if Status = Hired

Part (a)

(i) How many men were in the Donner party?

[2]:
```
Male = Donner[Sex==1,]
print(nrow(Male))
```

```
[1] 53
```

We mask the Donner dataframe and only lopok at values for which Sex = 1 (i.e. Males). Looking at the number of rows in the masked dataframe gives that that were 53 men in the Donner Party.

(ii) What was the survival rate for men?

[3]:
```
print(sum(Male$Survive)/nrow(Male))
```

```
[1] 0.4528302
```

If a man survived, their Surive variable is 1 - and if they didn't the Surve variable is 0. Adding up the entier Survive columsn, therefore, tells up how many total men survived. Dividing that by the total amount of men gives a survival rate of 45.28%.

Part (b)

[4]:
```
model_age = glm(Survive~Age, family=binomial)
summary(model_age)
```

```

Call:
glm(formula = Survive ~ Age, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.01255    0.37616   2.692  0.00711 **
Age         -0.03861    0.01505  -2.566  0.01030 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.86  on 87  degrees of freedom
Residual deviance: 113.41  on 86  degrees of freedom
AIC: 117.41

Number of Fisher Scoring iterations: 4
```

We simply take the coefficients provided in the coefficients table to get:

$$\ln(\text{odds}) = -0.03861 \cdot \text{Age} + 1.01255.$$

Part (c)

[5]:
```
model_sex = glm(Survive~Sex, family=binomial)
summary(model_sex)
```

```

Call:
glm(formula = Survive ~ Sex, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9163     0.3742   2.449   0.0143 *
Sex          -1.1055     0.4649  -2.378   0.0174 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.86  on 87  degrees of freedom
Residual deviance: 114.88  on 86  degrees of freedom
AIC: 118.88

Number of Fisher Scoring iterations: 4
```

Conducting a logistic regression with Sex instead of Age, we see that the P-value for Sex is slightly higher than for Age (0.0174 as compared to 0.0103) meaning Age is a better predictor.

Part (d)

[6]:
```
print(exp(confint(model_sex, parm="Sex")))
```

```
Waiting for profiling to be done…

    2.5 %    97.5 %
0.1287829 0.8061998
```

Creating a 95% confidence interval for the Sex parameter gives it in terms of log-odds, so we apply the exp function to get the odds ratio.

Part (e)

Note that the odds ratio is computed as $\frac{p}{1-p}$. The confidence interval is contained within the interval (0, 1) meaning that $p < 1 - p$. Since p is the probability that if Sex = 1 (male), then Survive = 1 (survives), this means that women in the Donner Party are more likely to survive than men.

Question 2

Part (a)

[7]:
```
model_sat = glm(Survive~Sex+Age+I(Age^2), family=binomial)
summary(model_sat)
```

```
Call:
glm(formula = Survive ~ Sex + Age + I(Age^2), family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.942540   0.592578   1.591   0.1117
Sex         -1.221738   0.511146  -2.390   0.0168 *
Age          0.073064   0.059619   1.226   0.2204
I(Age^2)    -0.002335   0.001346  -1.735   0.0828 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.86  on 87  degrees of freedom
Residual deviance: 103.72  on 84  degrees of freedom
AIC: 111.72

Number of Fisher Scoring iterations: 5
```

Again we look at the coefficients table to get:

$$\ln(\text{odds}) = 1.01255 - 1.221738 \cdot \text{Sex} + 0.073064 \cdot \text{Age} - 0.002335 \cdot \text{Age}^2$$

Part (b)

[8]:
```
model_reduced = glm(Survive~Sex+Age, family=binomial)
anova(model_reduced, model_sat, test="Chisq")
```

	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	Pr(>Chi) <dbl>
1	85	108.3333	NA	NA	NA
2	84	103.7247	1	4.608575	0.03181247

The p-value is 0.03, which means we reject the null hypothesis. There is a significant difference between the two models, so the Age^2 term is significant.

Part (c)

[9]:
```
presid = residuals(model_sat, type="pearson")
print(max(abs(presid)))
```

```
[1] 3.719524
```

We see that the maximum absolute value of the Pearson residual is 3.719 (which is positive as well in the original dataset)

Part (d)

[10]:
```
hired_table = table(Hired, Survive)
hired_table
```

```
     Survive
Hired  0  1
    0 30 44
    1  9  5
```

[11]:
```
chisq.test(hired_table)
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  hired_table
X-squared = 1.8137, df = 1, p-value = 0.1781
```

Conducting a χ^2 test on the contingency table, gives a p-value of 0.178, which means that there is nt a significant difference in the survival rates between hired and not hired emigrants in the Donner Party.

Question 3

Part (a)

[12]:
```
qnorm(1-0.03523/2)
```

```
2.10570483027573
```

To calculate the Z statistic, we first start by looking at the estimated coefficient for Change. Sinze it is positive, the z-value is also positive. We then look at its associated p-value. Since this is calculating at Pr(>|z|), we know that we are looking at twice the area in the right tail (since z is positive). Finally we apply the qnorm function (since this is a z-statistic the distribution is normal) with 1 - p/2 to calculate the z with the correct area to the left. This gives a z value of 2.106.

Part (b)

[13]:
```
42 + 1 - 3
```

```
40
```

The null deviance degrees of freedom is simply the number of data points, n, minus the number of variables in the null model (which is just 1 - MarriageBan). Therefore $n = 42 + 1 = 43$. To calculate the residual deviance degrees of freedom we subtract the number of variables in the model (in this case 3 - MarriageBan, Change, and UnionBan) which gives us that df $= 43 - 3 = \boxed{40}$

Part (c)

[14]:
```
UnionBan = 0
Change = 0.016
MarriageBan = exp(1.98+13.07*Change+18.12*UnionBan)/(1+exp(1.98+13.07*Change+18.
↪12*UnionBan))
print(MarriageBan)
```

```
[1] 0.8992682
```

We simply plug in the given values into the sigmoid funciton with the log-odds linear function given by the coefficients in the table. This gives us that the esitmated probability of MarriageBan in Oregon is 90%.

Question 4

Part (a)

The sign of Protein indicates that an increase in protein leads to a decrease in the predicted probability of Kids, or in other words, an increased probability that the cereal is marketed towards adults. In comparison with the other predictors (whose p-values are not significant to any reasonable level), the magnitude of the coefficient for Protein is extremely high indicating a strong relationship between Protein and Kids.

Part (b)

[15]:
```
pchisq(81.84, 69, lower.tail=FALSE)
```

```
0.138364120144726
```

To test the overall effectiveness of this model, we run a χ^2 test on the residual deviance with the given degrees of freedom (69). The p-value (calculated above) is 0.138 which is not significant to the 0.05 level. Therefore we fail to reject the null hypothesis and conclude that we can reject the claim that the model is effective (as comapred to the null model Kids~1).