MA 413 - Statistics for Data Science

Solutions to Exercise Sheet 12

1.
- Referring to pages 7-8 of Lecture 18 Slides, for the probit model we define:

$$\pi_i = \mathbb{E}[Y_i] \tag{1}$$

associated with covariate x_i, and we set

$$\pi_i = \phi(\beta_0 + \beta_1 x_i) \tag{2}$$

The argument can range over the entire line, but we know that $\phi(\beta_0 + \beta_1 x_i)$ ranges between 0 and 1.

- For the logit model we have:

$$\pi_i = \mathbb{E}[Y_i] \tag{3}$$

and

$$\log\{\frac{\pi_i}{1-\pi_i}\} = \beta_0 + \beta_1 x_i. \tag{4}$$

Having π_i range from 0 to 1, makes $g(\pi_i) = \frac{\pi_i}{1-\pi_i}$ range over the entire real line.

- When to use which?

Both models are abstractions of real-world situations, and both methods are theoretically similar. On the one hand, when it comes to practice logit models have pdfs which can be more easily integrated compared to the pdf of the probit model, which often needs simulation.

On the other hand, in practice we can sometimes see the popularity of one model taking place in a particular field. For insance:

Logit models are more widely used in health sciences like epidemiology partly because coefficients can be interpreted in terms of odds ratios.

Probit models can be used in some advanced econometric settings to account for non-constant error variances.

However, it is of importance to emphasize that both methods yield similar inferences.

2.
```
#Begin with ungrouped data
X_ungrouped <- c(0,0,0,0,1,1,1,1,2,2,2,2) #values of X
Y_ungrouped <- c(0,1,0,0,1,0,1,0,1,1,1,1) #trials from table


ungrouped_df <- data.frame(x=X_ungrouped, y=Y_ungrouped) #create dataframe


s1 <- factor(1:length(Y_ungrouped)) #vector of coefficient to be fitted for the saturated model
m0_ungrouped <- glm(formula = y ~ 1, family = binomial(link = logit), data = ungrouped_df) #M_0
m1_ungrouped <- glm(formula = y ~ x, family = binomial(link = logit), data = ungrouped_df) #M_1
ms_ungrouped <- glm(formula = y ~ s1, family = binomial(link = logit), data = ungrouped_df) #Saturated model

#Group the data so that it has binomials statistics to fit
X_binomial <- c(0,1,2)
proportion_success <- cbind(c(1,2,4), c(3,2,0)) #make the previous bernouilli from Y_ungrouped into a
    binomial summary statistics

s2 <- factor(1:3)
m0_grouped <- glm(formula = proportion_success~ 1, family = binomial(link = logit))
m1_grouped <- glm(formula = proportion_success~ X_binomial, family = binomial(link = logit))
ms_grouped <- glm(formula = proportion_success~ s2, family = binomial(link = logit))
```

(a) The loglikelihood are reported in Fig.1

```
> logLik(m0_ungrouped)
'log Lik.' -8.150319 (df=1)
> logLik(m1_ungrouped)
'log Lik.' -5.514129 (df=2)
> logLik(ms_ungrouped)
'log Lik.' -2.572005e-10 (df=12)
> logLik(m0_grouped)
'log Lik.' -4.972265 (df=1)
> logLik(m1_grouped)
'log Lik.' -2.336075 (df=2)
> logLik(ms_grouped)
'log Lik.' -1.843875 (df=3)
```

Figure 1: Loglikelihood from all models for grouped and ungrouped data

```
> summary(m0_ungrouped)

Call:
glm(formula = y ~ 1, family = binomial(link = logit), data = ungrouped_df)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.323  -1.323   1.038   1.038   1.038

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3365     0.5855   0.575    0.566

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16.301  on 11  degrees of freedom
Residual deviance: 16.301  on 11  degrees of freedom
AIC: 18.301

Number of Fisher Scoring iterations: 4

> summary(m0_grouped)

Call:
glm(formula = proportion_success ~ 1, family = binomial(link = logit))

Deviance Residuals:
      1       2       3
-1.3536  -0.3357   2.0765

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3365     0.5855   0.575    0.566

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.2568  on 2  degrees of freedom
Residual deviance: 6.2568  on 2  degrees of freedom
AIC: 11.945

Number of Fisher Scoring iterations: 4
```

Figure 2: Summary of M_0 for ungrouped and grouped data

(b) We report the summaries from M_0 and M_1, from both ungrouped and grouped data, respectively by figure Fig.2 and Fig.3. From those figures, we see that switching from a Bernouilli setting to a binomial statistics reduces deviance as well as the AIC, leading to a more well-fitting model.

Deviance residuals are overall close to 0 and getting closer to 0 when grouping the data, hence making our model less biased. This shows the importance of the form of the data entry on deviance. This can also be seen with loglikelihood results being bigger when grouping the data.

3. R code:

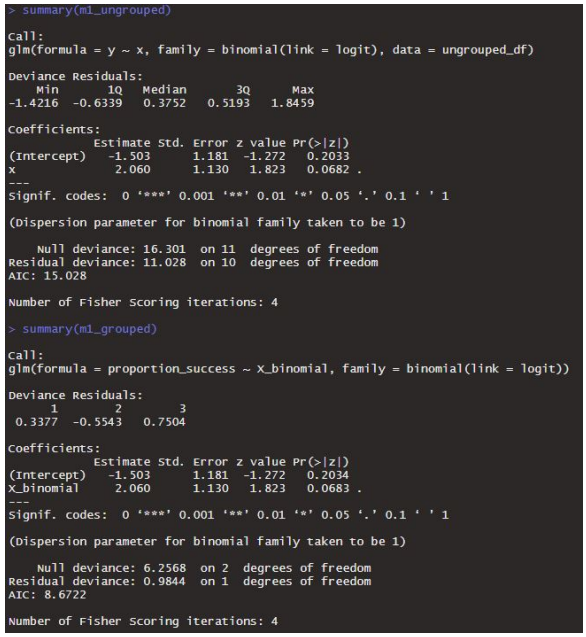
```
Y <- c(33, 38, 63, 108, 157, 159)
N <- c(3271, 2486, 7256, 8877, 5065, 3250)
Gender <- factor(rep(c("M", "F"), 3), levels=c("M", "F"))
Age <- c("<=13", "14-18", ">=19","<13", "14-18", ">=19")

missing <- data.frame(Y=Y, N=N, Gender=Gender, Age=Age)

fit <- glm( cbind(Y, N-Y) ~ Age, family=binomial, data=missing)

summary(fit)

Call:
glm(formula = cbind(Y, N - Y) ~ Age, family = binomial, data = missing)
```



```
> summary(m1_ungrouped)

Call:
glm(formula = y ~ x, family = binomial(link = logit), data = ungrouped_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4216  -0.6339   0.3752   0.5193   1.8459

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.503      1.181  -1.272   0.2033
x              2.060      1.130   1.823   0.0682 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16.301  on 11  degrees of freedom
Residual deviance: 11.028  on 10  degrees of freedom
AIC: 15.028

Number of Fisher Scoring iterations: 4

> summary(m1_grouped)

Call:
glm(formula = proportion_success ~ X_binomial, family = binomial(link = logit))

Deviance Residuals:
      1       2       3
 0.3377  -0.5543   0.7504

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.503      1.181  -1.272   0.2034
X_binomial     2.060      1.130   1.823   0.0683 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.2568  on 2  degrees of freedom
Residual deviance: 0.9844  on 1  degrees of freedom
AIC: 8.6722

Number of Fisher Scoring iterations: 4
```

Figure 3: Summary of M_1 for ungrouped and grouped data

```
Deviance Residuals:
     1       2       3       4       5       6
 0.000  -3.581  -8.350   0.000   2.251   9.429

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.5862     0.1750 -26.213  < 2e-16 ***
Age<13        0.1894     0.2000   0.947    0.344
Age>=19       0.7505     0.1877   4.000 6.34e-05 ***
Age14-18      0.9559     0.1894   5.047 4.49e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 237.77  on 5  degrees of freedom
Residual deviance: 176.53  on 2  degrees of freedom
AIC: 221.53

Number of Fisher Scoring iterations: 5
```

Figure 4: Output of the code for question 3

If we first have a look at the deviance residuals we see that we have large deviance residuals (see Fig.4). Columns 1-6 correspond to Min., 1st Qu., Median, Mean, 3rd Qu., Max, respectively. We can see that column 3 which corresponds to median deviance residual is not even close to zero, so this means that our model is biased in one direction (underestimated).

On the other hand, we should have $\frac{\text{residual deviance}}{\text{degrees of freedom}} \approx 1$, which is not the case for our model. Hence, this also implies that this is not a well-fitting model. This could be a result of, for example, overdispersion.

Another informative criterion which gives us clues about the well-fitting of the model is AIC-the Akaike Informtion Criterion. We know that a model with a low AIC has low complexity and a good fit. As it turns out, by no surprise, AIC is high in our model which once again implies that this model is not a well-fitting one.

4. (a) True. As seen in slide 10 from lecture 18, as any model we choose is a nested model of the saturated by just putting restriction on it, we can build a goodness-of-fit test using the following chi square distribution.

   Assuming the model truly holds, we have:

$$2\left(\ell_n\left(\hat{\beta}^S\right) - \ell_n\left(\hat{\beta}^M\right)\right) = D_M - D_S \xrightarrow{d} \chi^2_{p-q}$$

   where S, with p parameters, is the saturated model and M is the model from the question with q parameters ($q < p$). To further perform the test, one can use the p-value from a chi-square distribution with $n - p$ degrees of freedom by computing the probability to the right of the deviance value.

   (b) This happens when it is a complete separation. One can think of it in the following two ways. The likelihood based on $(x_1, y_1), \ldots, (x_n, y_n)$ is

$$\prod_{i=1}^{n} (\pi(x_i))^{y_i} (1 - \pi(x_i))^{(1-y_i)},$$

   and the log-likelihood is

$$\sum_{i=1}^{n} y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i)) = \sum_{i:x_i \leq 50} \log(\pi(x_i)) + \sum_{i:x_i > 50} \log(1 - \pi(x_i)),$$

   which is always non-positive when $\pi(x) \in (0, 1)$, and the best fit of $\pi(x)$ is the one such that the log-likelihood tends to zero. Given

$$\text{logit}(\pi(x)) = \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x,$$

   we obtain

$$\pi(x) = \frac{1}{1 + \exp(-\alpha - \beta x)}.$$

   Let $\alpha + \beta x = \alpha^* + \beta(x - 50)$, the MLE estimation does not exist, and the log-likelihood goes to zero as β goes to $-\infty$.

   On the other hand, the Newton-Raphson iteration or equivalently, the IRLS, fails to converge, and the fitted model goes to the true model as β goes to $-\infty$.