MTH6134 2022 Sample Exam

1. **Likelihood**

ESSAY 1.0 point 0.10 penalty editor

Suppose that $Y_i \sim N(\mu_i, \sigma^2)$ for $i = 1, 2, \ldots, n$ all independent with σ^2 known; and that $\mu_i = \beta_1 x_i + \beta_2 z_i$ where x_i and z_i are known covariates.

A) Write down the likelihood for the data $y_1, \ldots, y_n$.

B) Find the maximum likelihood estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ of β_1 and β_2.

C) Prove that $\hat{\beta}_1$ is an unbiased estimator of β_1.

D) Explain why $\hat{\beta}_1$ has a normal distribution.

Notes for grader:

- A) The likelihood is $L(\beta_1, \beta_2; y) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_1 x_i - \beta_2 z_i)^2\right)$. Students are not meant to proceed further expanding the quadratic expression in the exponent of the likelihood, but to use the exponential of the sum.
- B) Stemming from the log-likelihood $l(\beta_1, \beta_2; y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_1 x_i - \beta_2 z_i)^2$, students are meant to compute partial derivatives $\frac{\partial l}{\partial \beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^n x_i(y_i - \beta_1 x_i - \beta_2 z_i)$ and $\frac{\partial l}{\partial \beta_2} = \frac{1}{\sigma^2}\sum_{i=1}^n z_i(y_i - \beta_1 x_i - \beta_2 z_i)$. The maximum likelihood estimates satisfy the system $\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i z_i \\ \sum_{i=1}^n x_i z_i & \sum_{i=1}^n z_i^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n z_i y_i \end{pmatrix}$. The solution of this system are the maximum likelihood estimates $\hat{\beta}_1 = \frac{1}{\Delta}\left(\sum_{i=1}^n z_i^2 \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i z_i \sum_{i=1}^n z_i y_i\right)$ and $\hat{\beta}_2 = \frac{1}{\Delta}\left(\sum_{i=1}^n x_i^2 \sum_{i=1}^n z_i y_i - \sum_{i=1}^n x_i z_i \sum_{i=1}^n x_i y_i\right)$, with $\Delta = \sum_{i=1}^n x_i^2 \sum_{i=1}^n z_i^2 - \left(\sum_{i=1}^n x_i z_i\right)^2$. This is a standard calculation, albeit slightly lengthy.
- C) Recall that the maximum likelihood estimates are $\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i z_i \\ \sum_{i=1}^n x_i z_i & \sum_{i=1}^n z_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n z_i y_i \end{pmatrix}$, and that the expectation of Y_i is $E(Y_i) = \mu_i = \beta_1 x_i + \beta_2 z_i = (x_i \;\; z_i)\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$. Using the previous expectation in the vector $\begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n z_i y_i \end{pmatrix}$ above,

this vector is rewritten as

$$\begin{pmatrix} \sum_{i=1}^n x_i(x_i \quad z_i)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} \\ \sum_{i=1}^n z_i(x_i \quad z_i)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix}\end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i z_i \\ \sum_{i=1}^n x_i z_i & \sum_{i=1}^n z_i^2 \end{pmatrix}\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix}.$$

We collect what we have done and we have

$$E\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i z_i \\ \sum_{i=1}^n x_i z_i & \sum_{i=1}^n z_i^2 \end{pmatrix}^{-1}\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i z_i \\ \sum_{i=1}^n x_i z_i & \sum_{i=1}^n z_i^2 \end{pmatrix}\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix},$$

that is $E\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix}$. This is a way to simultaneously prove that both $\hat{\beta}_1, \hat{\beta}_2$ are unbiased estimates.

This can also be done individually by substituting the expectation of y_i in the explicit equation for $\hat{\beta}_1$. Let us work this second approach, noting that the expectation of Y_i is a vector product

$$E(\hat{\beta}_1) = \frac{1}{\Delta}\left(\sum_{i=1}^n z_i^2 \sum_{i=1}^n x_i(x_i \quad z_i)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} - \sum_{i=1}^n x_i z_i \sum_{i=1}^n z_i(x_i \quad z_i)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix}\right).$$

The part between brackets above becomes

$$(\sum_{i=1}^n z_i^2 \sum_{i=1}^n x_i^2 \quad \sum_{i=1}^n z_i^2 \sum_{i=1}^n x_i z_i)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix}$$

$-((\sum_{i=1}^n x_i z_i)^2 \quad \sum_{i=1}^n z_i^2 \sum_{i=1}^n x_i z_i)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix}$, which after simplification becomes

$E(\hat{\beta}_1) = \frac{1}{\Delta}\left(\sum_{i=1}^n z_i^2 \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i z_i)^2 \quad 0\right)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix}$ and then $E(\hat{\beta}_1) = \frac{1}{\Delta}(\sum_{i=1}^n z_i^2 \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i z_i)^2)\beta_1$. We finally note that the expression multiplying β_1 is precisely Δ and we have thus shown that $E(\hat{\beta}_1) = \beta_1$.

Either approach to show unbiasedness is equally fine.

- D) The distribution of $\hat{\beta}_1$ is normal because it is a linear combination of normal random variables. This can be seen by noting that $\hat{\beta}_1 = \sum_{i=1}^n c_i y_i$ with the c_i taken from the expression of $\hat{\beta}_1$. The explicit form of the c_i coefficients is

  $$c_i = \frac{1}{\Delta}\left((\sum_{j=1}^n z_j^2)x_i - (\sum_{j=1}^n x_j z_j)z_i\right).$$

  Note the change of summation indices to j to clarify the coefficients c_i. The argument based on the sum of normal random variables is enough.

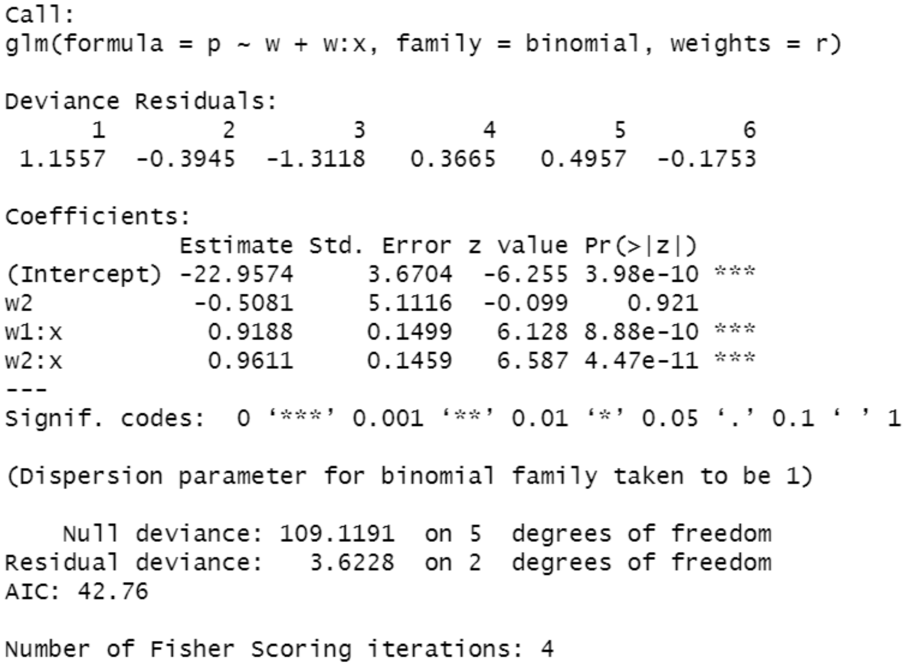
2. **Discharge data (1)**

ESSAY 1.0 point 0.10 penalty editor

The numbers of babies surviving to discharge from a hospital (y) out of the number admitted to neonatal intensive care (r) for two epochs (w) and three gestational ages (x) in weeks were recorded. Below are the data.

x	23	23	24	24	25	25
w	1	2	1	2	1	2
r	81	65	165	198	229	225
y	15	12	40	82	119	142

Let Y_{jk} denote the number of babies surviving to discharge out of the r_{jk} of gestational age x_k admitted to neonatal intensive care of epoch j. Then it is assumed that $Y_{jk} \sim \text{Bin}(r_{jk}, \pi_{jk})$ for $j = 1, 2$ and $k = 1, 2, 3$, all independent, where $\log(\pi_{jk}/(1 - \pi_{jk})) = \alpha_j + \beta_j x_k$. This model was fit to data using `R` and the following output was obtained:



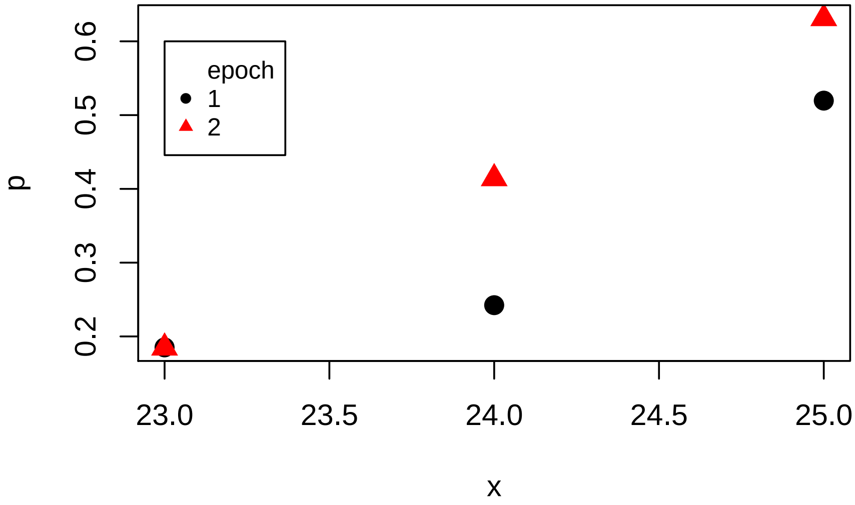
```
Call:
glm(formula = p ~ w + w:x, family = binomial, weights = r)

Deviance Residuals:
      1        2        3        4        5        6
 1.1557  -0.3945  -1.3118   0.3665   0.4957  -0.1753

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -22.9574     3.6704  -6.255 3.98e-10 ***
w2           -0.5081     5.1116  -0.099    0.921
w1:x          0.9188     0.1499   6.128 8.88e-10 ***
w2:x          0.9611     0.1459   6.587 4.47e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 109.1191  on 5  degrees of freedom
Residual deviance:   3.6228  on 2  degrees of freedom
AIC: 42.76

Number of Fisher Scoring iterations: 4
```

A) Plot the proportions of babies surviving to discharge against gestational age by epoch. What do you observe, what are your conclusions?

B) Write down the fitted logistic regression model for each epoch.

Notes for grader:

- A) Here is a plot of proportions of babies surviving discharge against gestational age by epoch.

  The plot suggests that the regression lines for the first two epochs are not parallel. In particular, most of the survival rates have improved from the first epoch to the second.

- B) The maximum likelihood estimates of $\alpha_1, \alpha_2, \beta_1$ and β_2 are $\hat{\alpha}_1 = -22.9574, \hat{\alpha}_2 = -23.4655, \hat{\beta}_1 = 0.9188$ and $\hat{\beta}_2 = 0.9611$. Hence the fitted regression model is

  $$\hat{\pi}_{1k} = 1/\left(1 + \exp(-(-22.9574 + 0.9188x_k))\right)$$

  for those babies at the first epoch, and

  $$\hat{\pi}_{2k} = 1/\left(1 + \exp(-(-23.4655 + 0.9611x_k))\right)$$

  for those at the second.
  The identity $e^u/(1+e^u) \equiv 1/(1+e^{-u})$ for the inverse logistic function was used above. In other words, the regression models can be equivalently written as

  $$\hat{\pi}_{1k} = e^{-22.9574+0.9188x_k}/(1 + e^{-22.9574+0.9188x_k})$$

and

$$\hat{\pi}_{2k} = e^{-23.4655+0.9611x_k}/(1 + e^{-23.4655+0.9611x_k}).$$

Either version is fine, provided the correct coefficients have been used.

3. **Discharge data (2)**

NUMERICAL 1.0 point 0.10 penalty

C) Write the value of the residual deviance.

- 3.6227808 ± 5e-2 ✓

4. **Discharge data (3)**

ESSAY 1.0 point 0.10 penalty editor

D) Use this deviance value to assess the goodness of fit of the model.

Notes for grader:

- D) Concerning the goodness of fit of the model, we are fitting a model with $p = 4$ parameters and the maximal model has $n = 6$ parameters. The data gives $D = 3.6228$, with p-value 0.1634 (computed with the χ^2_2 distribution). As the p-value is bigger than $\alpha = 0.05$, we have no evidence against the logistic regression model.

5. **Discharge data (4)**

NUMERICAL 1.0 point 0.10 penalty

E) Using the output of the model you just fit, estimate the difference $\beta_1 - \beta_2$.

- -0.0422486 ± 5e-2 ✓

6. **Discharge data (5)**

ESSAY 1.0 point 0.10 penalty editor

F) Give an approximate 95 percent confidence interval for $\beta_1 - \beta_2$.

G) Comment on the confidence interval.

Notes for grader:

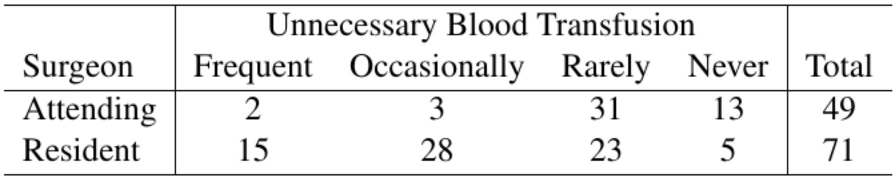
- F) An approximate 95% confidence interval for $\beta_1 - \beta_2$ is $\hat{\beta}_1 - \hat{\beta}_2 \pm 1.96\sqrt{\hat{v}^{33} + \hat{v}^{44}}$. The estimates give the difference $\hat{\beta}_1 - \hat{\beta}_2 = 0.9188 - 0.9611 = -0.0423$; the estimated variance of this difference is $\hat{v}^{33} + \hat{v}^{44} = 0.1499^2 + 0.1459^2 = 0.0438$ so that the standard deviation of the difference is $\sqrt{\hat{v}^{33} + \hat{v}^{44}} = 0.2093$. We collect all the information to retrieve the confidence interval $-0.0423 \pm 1.96 \cdot 0.2093 = -0.0423 \pm 0.4102$, that is $(-0.4525, 0.3679)$.

  G) As the interval contains zero, we cannot reject the null hypothesis of the equality between coefficients, i.e. $H_0 : \beta_1 = \beta_2$.

7. **Residents (1)**

ESSAY 1.0 point 0.10 penalty editor

A study of 49 attending physicians and 71 surgical residents in training at a university hospital was carried out to investigate whether the two groups of surgeons were applying unnecessary blood transfusions at different rates. For each surgeon, the number of blood transfusions prescribed unnecessarily in one year was recorded. The contingency table below summarizes the data.



	Unnecessary Blood Transfusion				
Surgeon	Frequent	Occasionally	Rarely	Never	Total
Attending	2	3	31	13	49
Resident	15	28	23	5	71

Let Y_{jk} denote the number of surgeons classified in row j and column k. Then it is assumed that the Y_{jk} for row j have a multinomial distribution with parameters $y_{j\cdot}$ and θ_{jk} for $j = 1, 2$ and $k = 1, 2, 3, 4$, and that the rows are independent, where $y_{j\cdot} = \sum_{k=1}^{4} y_{jk}$ and θ_{jk} is the probability that a surgeon is classified in row j and column k. The null hypothesis is that the distributions of unnecessary blood transfusions are the same for the two groups of surgeons.

A) Briefly explain how you would enter these data into `R`.

B) What command would you use to fit a log-linear model to the data?

Notes for grader:

- A) The data would be entered into `R` column by column.

```
c1<-c(2,15)
c2<-c(3,28)
c3<-c(31,23)
c4<-c(13,5)
y<-c(c1,c2,c3,c4)
```

Then the levels of the rown and column factors would be generated by

```
row<-gl(n=2,k = 1,length = 8)
column<-gl(n=4,k = 2, length=8)
```

Note that an equivalent formulation can be achieved entering the data row by row. The row and column factors need to be adapted.

- B) A log-linear model is fitted to the data using

```
blood<-glm(formula=y row+column,family = poisson)
```

Note that by default the link is `log` which can be written (or not) as part of the output.

8. **Residents (2)**

ESSAY 1.0 point 0.10 penalty editor

C) Obtain the expected values under the null hypothesis. Comment on what you observe.

Notes for grader:

- C) These values are obtained with the command `blood $fitted.values` that yield values 6.942, 10.058, 12.658, 18.342, 22.05, 31.95, 7.35, 10.65. These values have to be formatted into the shape of a table

	Unnecessary Blood Transfusion				
Surgeon	Frequent	Occasionally	Rarely	Never	Total
Attending	6.942	12.658	22.050	7.350	49
Resident	10.058	18.342	31.950	10.650	71

By comparing these values, we see that surgical residents in training applied unnecessary blood transfusions more frequently than attending physicians.

9. **Residents (3)**

CLOZE 0.10 penalty

D) Compute the deviance and write its value.

NUMERICAL 1 point

35.3312 ± 1e-1 ✓	

E) Compute the value of Pearson's goodness-of-fit statistic and write its value.

NUMERICAL 1 point

31.8814 ± 1e-1 ✓	

10. **Residents (4)**

ESSAY 1.0 point 0.10 penalty editor

F) What is your conclusion about the numbers of unnecessary blood transfusions for the two groups of surgeons.

Notes for grader:

- F) The critical value is $\chi^2_{3,0.05} = 7.8147$, which when compared against the statistics suggests that there is strong evidence that the distributions of unnecessary blood transfusions are not the same for the two groups of surgeons.

Total of marks: 11