

# MATH3821 Assignment 1

*Stephen Sung*

## Question 1

For Simple Linear Regression model  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$  where  $\epsilon_i \sim N(0, \sigma^2)$ .

- a) Let  $\beta_0 = \alpha - \beta_1 \bar{x}$ . Then the SLR model can be expressed as  $y_i = \alpha + \beta_1(x_i - \bar{x}) + \epsilon_i$ .
- b)  $\alpha$  is the intercept of the new model?
- c) To find the closed form formula of the LSE,

$$RSS(\beta_1) = \sum_{i=1}^n [y_i - (\alpha + \beta_1(x_i - \bar{x}))]^2$$
$$\frac{dRSS(\beta_1)}{d\alpha} = -2 \sum_{i=1}^n (y_i - (\alpha + \beta_1(x_i - \bar{x}))) \quad (1)$$

$$\frac{dRSS(\beta_1)}{d\beta_1} = -2 \sum_{i=1}^n (y_i - (\alpha + \beta_1(x_i - \bar{x}))(x_i - \bar{x})) \quad (2)$$

Let Equation (1) = 0

$$\begin{aligned} -2 \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}_1(x_i - \bar{x}))) &= 0 \\ \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\alpha}_i - \sum_{i=1}^n \hat{\beta}_1 x_i + \sum_{i=1}^n \hat{\beta}_1 \bar{x} &= 0 \\ n\hat{\alpha}_i = \sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i + n\hat{\beta}_1 \bar{x} \\ \hat{\alpha}_i &= \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} \\ \hat{\alpha}_i &= \bar{y} \end{aligned}$$

Let Equation (2) = 0

$$\begin{aligned} -2 \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}_1(x_i - \bar{x}))(x_i - \bar{x})) &= 0 \\ \sum_{i=1}^n y_i(x_i - \bar{x}) - \sum_{i=1}^n \hat{\alpha}(x_i - \bar{x}) - \sum_{i=1}^n \hat{\beta}_1(x_i - \bar{x})^2 &= 0 \\ \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (y_i - \hat{\alpha})(x_i - \bar{x}) \end{aligned}$$

since  $\hat{\alpha} = \bar{y}$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

d)

$$\begin{aligned} Var[\hat{\alpha}] &= Var\left[\frac{1}{n} \sum_{i=1}^n y_i\right] \\ &= \frac{1}{n^2} Var\left[\sum_{i=1}^n y_i\right] \end{aligned}$$

Since all  $y_i$ 's are uncorrelated

$$\begin{aligned} &= \frac{1}{n^2} \sum_{i=1}^n Var[y_i] \\ &= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 \\ &= \frac{n\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Therefore  $Var[\hat{\alpha}] = \frac{\sigma^2}{n}$ .

We note that  $\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})y_i$ , since

$$\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^n y_i(x_i - \bar{x}) - \sum_{i=1}^n \bar{y}(x_i - \bar{x})$$

and we notice that

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \frac{1}{n}n \sum_{i=1}^n x_i = 0$$

To calculate  $Var[\hat{\beta}_1]$ ,

$$\begin{aligned} Var[\hat{\beta}_1] &= Var\left[\frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\ &= \frac{1}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} Var\left[\sum_{i=1}^n (x_i - \bar{x})y_i\right] \\ &= \frac{1}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} \sum_{i=1}^n Var[(x_i - \bar{x})y_i] \\ &= \frac{1}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} \sum_{i=1}^n (x_i - \bar{x})^2 Var[y_i] \\ &= \frac{1}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Let  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$  and  $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})y_i$ .

To calculate  $Cov[\hat{\alpha}, \hat{\beta}_1]$

$$\begin{aligned}
Cov[\hat{\alpha}, \hat{\beta}_1] &= Cov[\bar{y}, \hat{\beta}_1] \\
&= Cov\left[\bar{y}, \frac{S_{xy}}{S_{xx}}\right] \\
&= \frac{1}{S_{xx}} Cov[\bar{y}, S_{xy}] \\
&= \frac{1}{S_{xx}} Cov\left[\frac{1}{n} \sum_{i=1}^n y_i, S_{xy}\right] \\
&= \frac{1}{nS_{xx}} Cov\left[\sum_{i=1}^n y_i, \sum_{j=1}^n (x_j - \bar{x})y_j\right] \\
&= \frac{1}{nS_{xx}} \sum_{i=1}^n \sum_{j=1}^n (x_j - \bar{x}) Cov[y_i, y_j]
\end{aligned}$$

When  $i \neq j$ ,  $Cov[y_i, y_j] = 0$  since all  $y_i$  are uncorrelated with each other, and  $Cov[y_i, y_j] = Var[y_i]$  when  $i = j$

$$\begin{aligned}
&= \frac{1}{nS_{xx}} \sum_{i=1}^n (x_i - \bar{x}) \sigma^2 \\
&= 0
\end{aligned}$$

e)

```

set.seed(1234567)
x = runif(1000)
eps = rnorm(1000)
y = 5 + 10*x + eps
model <- y~x

# nablaRSS <- function(b) c(-2 * sum(Sales - b[1] - b[2] * TV), -2 * sum((Sales - b[1] - b[2] * TV) * TV) * T
# bn <- c(0, 0)
# gamma <- 0.0000001 # step size parameter
# kmax <- 1000000 ; for (k in 0:kmax) {
#   bnp1 <- bn - gamma * nablaRSS(bn)
#   if (k %% 100000 == 0)
#     { cat("b: ", bnp1, " -- RSS: ", RSS(bnp1[1], bnp1[2]), "\n") }
#   bn <- bnp1
# }

#Gradient Descent
RSS <- function(b) c(-2 * sum(y - b[1] - b[2] * x), -2 * sum((y - b[1] - b[2] * x) * x))
bn <- c(0,0)
gamma <- 0.00001
kmax <- 1000000
for (k in 0:kmax) {
  bnp1 <- bn - gamma * RSS(bn)
  bn <- bnp1
}

```

f)

g)

## Question 2

Given  $n$  independent binary random variables  $Y_1 \cdots Y_n$  with

$$P(Y_i = 1) = \pi_i \text{ and } P(Y_i = 0) = 1 - \pi_i$$

The probability function of  $Y_i$  is:

$$\pi_i^{Y_i} (1 - \pi_i)^{1-Y_i}$$

where  $Y_i = 0$  or  $1$

- a) For a probability function to belong to the exponential family of distributions, it must follow the formula:

$$f(y; \theta, \phi) = K(y, \frac{p}{\phi}) \exp \left( \frac{p}{\phi} \{y\theta - c(\theta)\} \right)$$

For the given probability density function:

$$\begin{aligned} f(y; \pi) &= \pi_i^y (1 - \pi_i)^{1-y} \\ &= \exp (\log \pi_i^y (1 - \pi_i)^{1-y}) \\ &= \exp (\log \pi_i^y + \log(1 - \pi_i)^{1-y}) \\ &= \exp (y \log \pi_i + (1 - y) \log(1 - \pi_i)) \\ &= \exp \left( y \log \left( \frac{\pi}{1 - \pi} \right) + \log(1 - \pi) \right) \end{aligned}$$

With  $p = 1$  and  $\phi = 1$ , the above equation follows the form of the exponential family of distribution where  $K(y, \frac{p}{\phi}) = 1$ ,  $\theta = \log(\frac{\pi}{1-\pi})$  and  $c(\theta) = -\log(1 - \pi) = -\log(1 - \frac{e^\theta}{1+e^\theta})$  where  $\pi = \frac{e^\theta}{1+e^\theta}$ .

- b) As seen in 2a, the naturalised parameter is  $\theta = \log(\frac{\pi}{1-\pi})$
- c) As seen in 2a, the cumulant generator is  $c(\theta) = -\log(1 - \frac{e^\theta}{1+e^\theta})$ . Since  $E[Y] = c'(\theta)$ ,  $c'(\theta) = -(\frac{e^\theta}{1+e^\theta}) = -(-\pi) = \pi$ . Therefore,  $E[Y] = \pi$ .
- d) Given the link function:

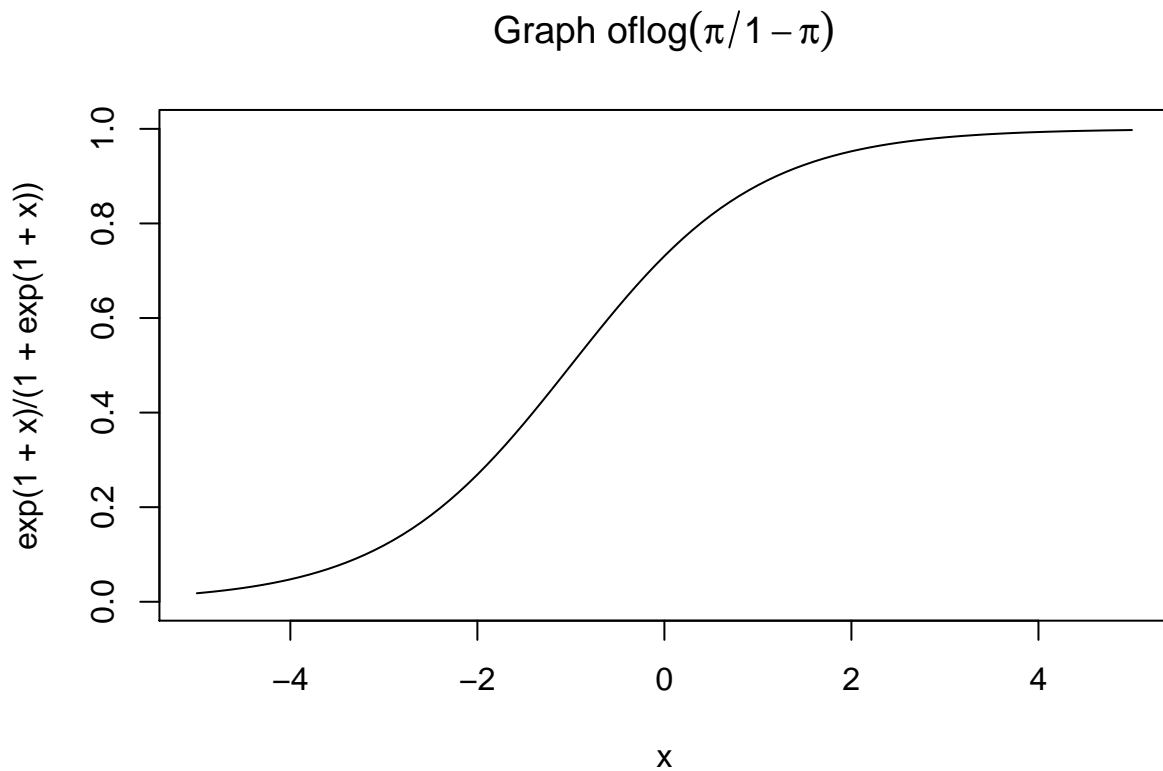
$$g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = e^{x^T \beta}$$

it can be rearranged in terms of the probability  $\pi$ ,

$$\begin{aligned} e^{x^T \beta} &= \log\left(\frac{\pi}{1 - \pi}\right) \\ e^{x^T \beta} - \pi e^{x^T \beta} &= \pi \\ \pi &= \frac{e^{x^T \beta}}{1 + e^{x^T \beta}} \end{aligned}$$

- e)

```
curve(exp(1+x)/(1+exp(1+x)), xlim = c(-5, 5), ylim = c(0, 1), main=expression(paste("Graph of", log(pi/
```



It shows the log odds of the insecticide working with a given probability?

f) The following probability density function:

$$f(y; \theta) = \frac{1}{\phi} \exp \left( \frac{(y - \theta)}{\phi} - \exp \left[ \frac{(y - \theta)}{\phi} \right] \right)$$

is NOT in the exponential family of distributions as it does not follow the form of a probability density function in the exponential family.

### Question 3

a)

```
titanic <- read.table('titanic.txt', header=TRUE)
head(titanic)
```

```
##              Name PClass   Age   Sex
## 1      Allen, Miss Elisabeth Walton    1st 29.00 female
## 2      Allison, Miss Helen Loraine     1st  2.00 female
## 3      Allison, Mr Hudson Joshua Creighton    1st 30.00   male
## 4 Allison, Mrs Hudson JC (Bessie Waldo Daniels)    1st 25.00 female
## 5      Allison, Master Hudson Trevor     1st  0.92   male
## 6      Anderson, Mr Harry              1st 47.00   male
## Survived
## 1      1
## 2      0
```

```
## 3      0
## 4      0
## 5      1
## 6      1
```

```
summary(titanic)
```

```
##              Name      PClass      Age      Sex
## Carlsson, Mr Frans Olof    : 2   1st:226   Min.   : 0.17   female:288
## Connolly, Miss Kate       : 2   2nd:212   1st Qu.:21.00   male  :468
## Kelly, Mr James           : 2   3rd:318   Median :28.00
## Abbing, Mr Anthony        : 1                Mean  :30.40
## Abbott, Master Eugene Joseph: 1                3rd Qu.:39.00
## Abbott, Mr Rossmore Edward : 1                Max.   :71.00
## (Other)                   :747
##      Survived
## Min.   :0.000
## 1st Qu.:0.000
## Median :0.000
## Mean   :0.414
## 3rd Qu.:1.000
## Max.   :1.000
##
```

b)

```
attach(titanic)
table(titanic$Sex)
```

```
##
## female    male
##      288    468
```

```
tapply(titanic$Survived,titanic$Sex,mean)
```

```
##      female      male
## 0.7534722 0.2051282
```

```
summary(lm(titanic$Survived~titanic$Sex))
```

```
##
## Call:
## lm(formula = titanic$Survived ~ titanic$Sex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.7535 -0.2051 -0.2051  0.2465  0.7949
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.75347    0.02445   30.82  <2e-16 ***
## titanic$Sexmale -0.54834    0.03107  -17.65  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4149 on 754 degrees of freedom
## Multiple R-squared:  0.2923, Adjusted R-squared:  0.2913
```

```
## F-statistic: 311.4 on 1 and 754 DF, p-value: < 2.2e-16
```

c)

```
titanic.glm <- glm(titanic$Survived~titanic$Age,family=binomial('logit'))
summary(titanic.glm)
```

```
##
## Call:
## glm(formula = titanic$Survived ~ titanic$Age, family = binomial("logit"))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.1418  -1.0489  -0.9792   1.3039   1.4801
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.081428   0.173862  -0.468   0.6395
## titanic$Age -0.008795   0.005232  -1.681   0.0928 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1025.6  on 755  degrees of freedom
## Residual deviance: 1022.7  on 754  degrees of freedom
## AIC: 1026.7
##
## Number of Fisher Scoring iterations: 4
```

```
exp(titanic.glm$coefficients[2])
```

```
## titanic$Age
##      0.9912439
```

```
exp(titanic.glm$coefficients[2])
```

```
## titanic$Age
##      0.9912439
```

d)

e)

f)

g)

h)

i)