

ass3

2022-11-23

### Question 1

a)

We have

$$\begin{aligned}\log f(y; r, \lambda_i) &= \log y^{r-1} + \log e^{-y/\lambda_i} - \log \lambda_i^r - \log \Gamma(x) \\ &= (r-1) \log y - \frac{y}{\lambda_i} - r \log \lambda_i - \log \Gamma(x) \\ &= (r-1) \log y - yr\theta_i - r \log \lambda_i - \log \Gamma(x) \\ &= (r-1) \log y - yr\theta_i - r \log ((r\theta_i)^{-1}) - \log \Gamma(x) \\ &= (r-1) \log y - yr\theta_i + r \log(r\theta_i) - \log \Gamma(x) \\ &= (r-1) \log y - yr\theta_i + r \log(r) + r \log(\theta_i) - \log \Gamma(x) \\ &= -yr\theta_i + r \log \theta_i + r \log(r) + (r-1) \log y - \log \Gamma(x) \\ &= -r(y\theta_i - \log \theta_i) + r \log(r) + (r-1) \log y - \log \Gamma(x) \\ &= \frac{y\theta_i - \log \theta_i}{-1/r} + r \log(r) + (r-1) \log y - \log \Gamma(x) \\ &= \frac{y\theta_i - \log \theta_i}{-1/r} + (r \log r + (r-1) \log y - \log \Gamma(r)).\end{aligned}$$

Thus, it follows that  $f(y; r, \lambda_i)$  belongs to the exponential family with  $b(\theta_i) = \log \theta_i$ ,  $\phi/A_i = -1/r$  with  $\phi = 1/r$  and  $A_i = -1$ .

b)

We have

$$b(\theta_i) = \log \theta_i$$

so  $E Y_i = b'(\theta_i) = \frac{1}{\theta_i} = r\lambda_i = \mu_i$ . Similarly,

$$\text{Var}(Y_i) = b''(\theta_i)\phi/A_i = \frac{-1}{\theta_i^2} \frac{-1}{r} = \frac{1}{\theta_i^2 r} = \frac{r^2 \lambda_i^2}{r} = r\lambda_i^2.$$

Furthermore, the CLF  $g(\mu_i) = (b')^{-1}(\mu_i) = \frac{1}{\mu_i}$ .

c)

We have the Pearson's statistic

$$\begin{aligned}
P &= \phi \sum_{i=1}^n \frac{(Y_i - E_{\hat{\beta}} Y_i)^2}{\text{Var}_{\hat{\beta}}(Y_i)} \\
&= \frac{1}{r} \sum_{i=1}^n \frac{\left(Y_i - \frac{1}{\hat{\theta}_i}\right)^2}{1/(\hat{\theta}_i^2 r)} \\
&= \frac{1}{r} \sum_{i=1}^n \frac{\left(\frac{Y_i \hat{\theta}_i}{\hat{\theta}_i} - \frac{1}{\hat{\theta}_i}\right)^2}{1/(\hat{\theta}_i^2 r)} \\
&= \frac{1}{r} \sum_{i=1}^n \frac{\left(Y_i \hat{\theta}_i - 1\right)^2}{\hat{\theta}_i^2 / (\hat{\theta}_i^2 r)} \\
&= \frac{1}{r} \sum_{i=1}^n \frac{(Y_i \hat{\theta}_i)^2 - 2Y_i \hat{\theta}_i + 1}{\hat{\theta}_i^2 / (\hat{\theta}_i^2 r)} \\
&= \sum_{i=1}^n \left( (Y_i \hat{\theta}_i)^2 - 2Y_i \hat{\theta}_i + 1 \right).
\end{aligned}$$

Furthermore,  $\tilde{\theta}_i = (b')^{-1}(Y_i) = 1/Y_i$  then

$$\begin{aligned}
D &= 2 \sum_{i=1}^n A_i (Y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)) \\
&= -2 \sum_{i=1}^n \left( Y_i \left( \frac{1}{Y_i} - \hat{\theta}_i \right) - \log \frac{1}{Y_i} + \log \hat{\theta}_i \right) \\
&= -2 \sum_{i=1}^n \left( Y_i \left( \frac{1}{Y_i} - \hat{\theta}_i \right) + \log Y_i + \log \hat{\theta}_i \right) \\
&= -2 \sum_{i=1}^n (1 - Y_i \hat{\theta}_i + \log Y_i + \log \hat{\theta}_i).
\end{aligned}$$

Finally, for the working matrix  $\hat{W}$ ,

$$\begin{aligned}
\hat{w}_{ii} &= \frac{A_i}{[g'(\mu_i)]^2 b''(\theta_i)} \\
&= \frac{1}{\frac{1}{\hat{\mu}_i^4} \frac{-1}{\hat{\theta}^2}} \\
&= \frac{1}{\frac{\hat{\theta}_i^4 - 1}{\hat{\theta}^2}} \\
&= \frac{-1}{\hat{\theta}_i^2}.
\end{aligned}$$

## Question 2

a)

There are 18 observations. The estimated parameters  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (-0.021, 0.017, 0.01)$ . Similarly,  $\tilde{\phi} = \frac{D_{\Omega}}{n-p-1} = \frac{0.3}{18-2-1} = 0.02$ . To test for  $H_0 : \beta_1 = 0$  against  $H_1 : \beta_1 \neq 0$ , we compute  $t_1 = \frac{\hat{\beta}_1}{\sqrt{(\hat{I}_F^{-1})_{11}}} =$

$\frac{0.017}{0.001} = 17$  where  $\left(\hat{I}_F^{-1}\right)_{11}$  is the square of the standard error of the first variable. Since the test statistic  $t_1 = 17$  is larger than  $t_{15;1-0.05/2} = 2.13$ , we reject  $H_0$ . The 96% confidence interval for  $\beta_1$  is

$$\hat{\beta}_1 \pm t_{15;1-0.04/2} \sqrt{\left(\hat{I}_F^{-1}\right)_{11}} = 0.017 \pm 2.25 \times 0.01 = [-0.0055, 0.0395].$$

b)

To investigate the relevance of the variables  $X1$  and  $X2$ , it suffices to look at the corresponding  $p$ -values in the output table of the command `summary(model)`. This shows that the  $p$ -value of  $X2$  is smaller than 0.05 so we can not dismiss the relevance of  $X2$ . Similarly, we test for the relevance of  $X1$  in a) and rejected the null hypothesis that  $\beta_1 = 0$  so  $X1$  is relevant.

c)

If not given, the standard errors in the first table could have been recovered as the square root of the variance of the sample. If the estimates are given, then the  $t$ -values could be (approximately) recovered by the formula

$$t_i = \frac{\hat{\beta}_i}{\sqrt{\left(\hat{I}_F^{-1}\right)_{ii}}}$$

where  $\sqrt{\left(\hat{I}_F^{-1}\right)_{ii}}$  is the standard error of the  $i$  predictor. It follows that the  $p$ -values could be recovered from the test statistic and the degree of freedom. Similarly, if the  $t$ -values are given, the estimates could be recovered.

For the second table, the degree of freedom can be recovered from the number of observations and the number of predictors. If the residuals deviance are given then we could calculate the deviances and the  $p$ -values since the test statistic will also be known. On the other hand, we are given the residual deviances for the variables but not the NULL one, then we could not calculate the deviance of the first variable since it depends on the residual deviance of the NULL one. Furthermore, if the deviances are known then we could not compute the residual one because the system of equation will be underdetermined. Also, by the result of Exercise 1, the residual deviance  $D$  of the models can be determined by the responses  $Y_i$  and the fitted canonical parameters  $\hat{\theta}_i$  but since  $\hat{\theta}_i = x_i^T \hat{\beta}$ , we could recover  $\hat{\theta}_i$  and therefore, the deviances of the models. And if the deviances are known, everything in the second table can be recovered.

d)

We use that option because the dispersion parameter  $\phi$  is not known beforehand. It is not possible to derive what we would have obtained if we used the Chi-square test option because the Chisquare test assumes that  $\phi$  is known and the test statistics use  $\phi$  instead of the estimated one in the  $F$  test. ### e) Since we have the dispersion parameter  $\hat{\phi}$ , the Pearson statistic  $P$  can be derived as

$$P = \hat{\phi}(n - p - 1) = 0.001958 * (18 - 2 - 1) = 0.0029.$$

It is not possible to derive  $\hat{W}$  because as we have shown in exercise 1, the element  $\hat{w}_{ij}$  of the working matrix  $\hat{W}$  depends on the estimated  $\hat{\theta}_i$  which in turn depends on the input  $x_i$ . Since we do not know  $x_i$  (not given in the question), we can not derive  $\hat{w}_{ij}$ .

### Question 3

a)

```
data3=read.table("psi.txt",header=TRUE)
attach(data3)
```

The following table gives us some insight about the dataset. Note that the students that have not been instructed with the PSI method are, a priori, more expected to fail. This observation is supported by the difference in the passing rate of both groups. However, we still need more information to draw any conclusions.

```
n00 = length(data3$gpa[data3$passed == 0 & data3$psi == 0])
n01 = length(data3$gpa[data3$passed == 0 & data3$psi == 1])
n10 = length(data3$gpa[data3$passed == 1 & data3$psi == 0])
n11 = length(data3$gpa[data3$passed == 1 & data3$psi == 1])
```

```
percent0 = n10 / (n00 + n10); percent0
```

```
## [1] 0.1666667
```

```
percent1 = n11 / (n01 + n11); percent1
```

```
## [1] 0.5714286
```

```
table_data = matrix(c(n00,n01,n10,n11), ncol=2, byrow=TRUE)
colnames(table_data) = c('No PSI','PSI')
rownames(table_data) = c('Failed','Passed')
table=as.table(table_data)
table
```

```
##           No PSI PSI
## Failed      15   6
## Passed       3   8
```

The table below shows the mean of the students for each of the combinations of the binary variables 'passed' and 'psi'. The only significant difference is that the mean of the students that passed and did not have PSI is higher than the average grade of the students that passed and did have PSI. However, this could be due to the fact that only 3 students passed. Probably, only the outstanding students managed.

```
m00 = mean(data3$gpa[data3$passed == 0 & data3$psi == 0]);
m01 = mean(data3$gpa[data3$passed == 0 & data3$psi == 1]);
m10 = mean(data3$gpa[data3$passed == 1 & data3$psi == 0]);
m11 = mean(data3$gpa[data3$passed == 1 & data3$psi == 1]);
```

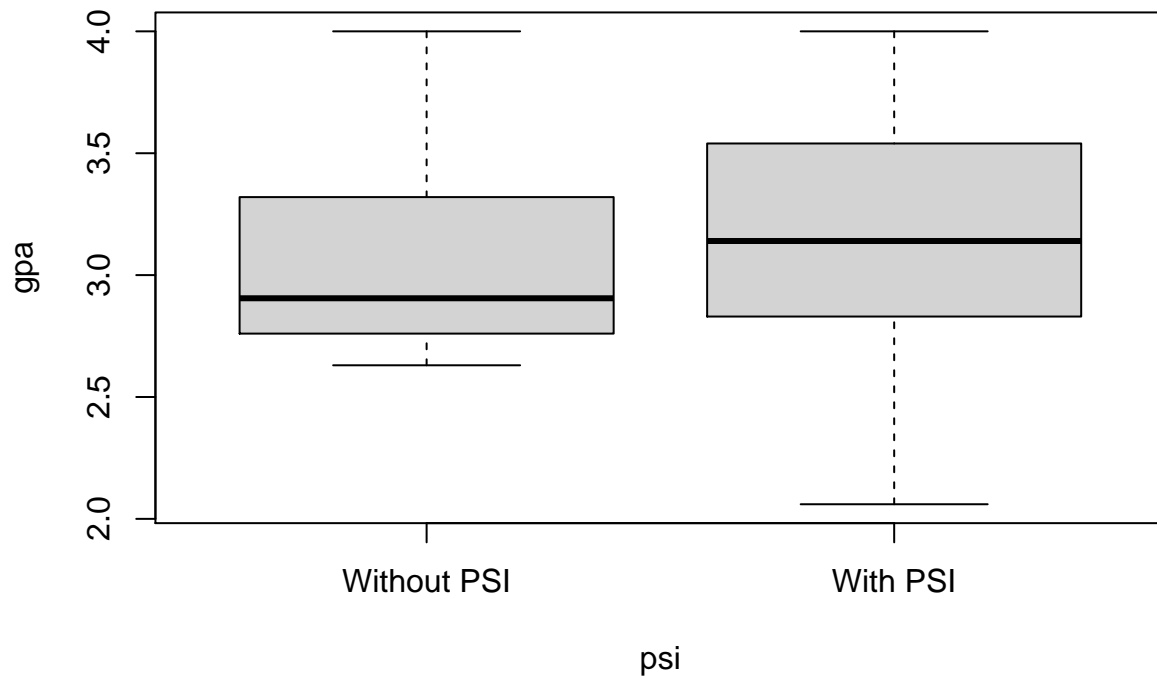
```
table_data2 = matrix(c(m00,m01,m10,m11), ncol=2, byrow=TRUE);
colnames(table_data2) = c('No PSI','PSI');
rownames(table_data2) = c('Failed','Passed');
table2=as.table(table_data2);
table2
```

```
##           No PSI      PSI
## Failed 2.976000 2.891667
## Passed 3.726667 3.322500
```

The following will give us more insight about the data. Note how the students that did have PSI have a higher average and also more variance. The plot suggests a better overall performance of the students with PSI.

```
boxplot(gpa~psi,data=data3,names=c("Without PSI","With PSI"), main='Box-plot of the average grade')
```

## Box-plot of the average grade



Now, we can try different models and compare how they fit, after converting the variable 'psi' into a factor:

```
data3$psi = as.factor(data3$psi)
```

```
model3_1=glm(passed~psi*gpa,data=data3,family=binomial)
anova(model3_1, test = 'Chisq')
```

```
## Analysis of Deviance Table
```

```
##
```

```
## Model: binomial, link: logit
```

```
##
```

```
## Response: passed
```

```
##
```

```
## Terms added sequentially (first to last)
```

```
##
```

```
##
```

```
##          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
```

```
## NULL                31      41.183
```

```
## psi                 1    5.8418      30      35.342 0.015650 *
```

```
## gpa                 1    9.0885      29      26.253 0.002572 **
```

```
## psi:gpa            1    1.8725      28      24.381 0.171189
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model3_2=glm(passed~psi+gpa,data=data3,family=binomial)
```

```
anova(model3_2, test = 'Chisq')
```

```
## Analysis of Deviance Table
```

```
##
## Model: binomial, link: logit
##
## Response: passed
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                31      41.183
## psi   1   5.8418      30      35.342 0.015650 *
## gpa   1   9.0885      29      26.253 0.002572 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first of them does not seem to be a trustful one, for the it has a non significant  $p$ -value for the interaction term of the variables 'psi' and 'passed'. We drop that term for the second model and see that both of them are significant. Thus, that would be a good model. However, if we only consider the interaction term, as we do in the third of the models, we achieve greater levels of significance, for the  $p$ -value is several orders of magnitude smaller than the  $p$ -values achieved by the second model. To have a second 'opinion', we can also check their summaries:

```
summary(model3_1)
```

```
##
## Call:
## glm(formula = passed ~ psi * gpa, family = binomial, data = data3)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6292  -0.5658  -0.1317   0.6220   2.1098
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -20.632     10.397  -1.984   0.0472 *
## psi1         15.085     11.181   1.349   0.1773
## gpa           5.681      2.970   1.913   0.0558 .
## psi1:gpa     -3.811      3.245  -1.174   0.2403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 41.183  on 31  degrees of freedom
## Residual deviance: 24.381  on 28  degrees of freedom
## AIC: 32.381
##
## Number of Fisher Scoring iterations: 6
```

```
summary(model3_2)
```

```
##
## Call:
## glm(formula = passed ~ psi + gpa, family = binomial, data = data3)
##
## Deviance Residuals:
```

```
##      Min      1Q   Median      3Q      Max
## -1.8396 -0.6282 -0.3045   0.5629   2.0378
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -11.602      4.213  -2.754  0.00589 **
## psi1          2.338      1.041   2.246  0.02470 *
## gpa           3.063      1.223   2.505  0.01224 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 41.183  on 31  degrees of freedom
## Residual deviance: 26.253  on 29  degrees of freedom
## AIC: 32.253
##
## Number of Fisher Scoring iterations: 5
```

Note that the  $p$ -values from the first model are much worse –i.e. larger– than the ones for the second model and, similarly, the ones for the second model are larger than the  $p$ -values for the third model. Before making a decision, we can also check the AIC of the three models:

```
AIC(model3_1, model3_2)
```

```
##      df      AIC
## model3_1  4 32.38065
## model3_2  3 32.25315
```

Again, the AIC confirms our thoughts: the best model seems to be the third one. So from now on, we will stick to that model.

Seeing the results of the model, which are significant enough,