Answers logistic regression, day 1

Exercise 1

```
library(readr)
ova <- read_csv("ova.csv", show_col_types = FALSE)
View(ova)
```

a Make 2 by 2 table

```
tab <- xtabs(~figo+death, data=ova)
tab
```

```
##     death
## figo   0   1
##    0  98 164
##    1  14  82
```

b. The chi-square test is most suited to compare the probability to die within 4 years between the two figo stages.

```
# calculate row proportions
proptab <- proportions(tab, "figo")
proptab
```

```
##     death
## figo         0         1
##    0 0.3740458 0.6259542
##    1 0.1458333 0.8541667
```

```
# and perform a chi-square test
chi2test <- chisq.test(tab)
chi2test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 15.976, df = 1, p-value = 6.416e-05
```

In the group with Figo=1 the probability to die within four years is 0.854, in the group with Figo=0 the probability is 0.626. There is a statistically significant association between figo staging and death (p = 6^{-5}). c. The odds ratio is (p1/(1-p1))/(p0/(1-p0)) = (0.854/(1-0.854))/(0.626/(1-0.626) = 3.5.

d. Perform a logistic regression analysis with death as dependent variable and figo as independent variable.

```
model.lr1 <- glm(death~figo, family=binomial, data=ova)
summary(model.lr1 )
```

```
##
## Call:
## glm(formula = death ~ figo, family = binomial, data = ova)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5149     0.1277   4.033 5.51e-05 ***
## figo          1.2528     0.3161   3.963 7.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 444.89  on 357  degrees of freedom
## Residual deviance: 426.16  on 356  degrees of freedom
## AIC: 430.16
##
## Number of Fisher Scoring iterations: 4
```

d. We see that the p-value for figo is very small, indicating that there is a significant association between figo state and the risk of dying. This is the same conclusion as the conclusion from the chi-squared test.

The regression coefficient for figo is 1.253. The corresponding oddsratio is exp(1.253) = 3.5. Exactly the same result as in 1.b.

e. The model on the logit scale is $log\left(\frac{\pi}{1-\pi}\right)$ = logit(P(death)) = 0.515 +1.253* figo. On the probability scale the model is P(death) = exp(0.515 +1.253* figo) / (1+ exp(0.515 +1.253* figo)). For someone with figo = 0, the predicted probability to die within 4 years is exp(0.515) / (1+ exp(0.515)) = 0.626. For someone with figo=1, P(death) = exp(0.515 +1.253) / (1+ exp(0.515 +1.253)) = 0.854. The results are the same as calculated in 1.b.

f. The 95% confidence interval for beta is (b-1.96 se(b), b+ 1.96se(b)). This gives (1.253 - 1.96$\times$ 0.316 , 1.253 + 1.96$\times$ 0.316) = (0.633, 1.872). The 95% CI for the oddsratio for figo=1 versus figo=0 is (exp(0.633 , exp(1.872))= (1.884, 6.503)

Check calculation of confidence interval

```
# for log OR
confint.default(model.lr1)
```

```
##                 2.5 %   97.5 %
## (Intercept) 0.2646549 0.765143
## figo        0.6331994 1.872327
```

```
# for OR
exp(confint.default(model.lr1))
```

```
##                2.5 %   97.5 %
## (Intercept) 1.302981 2.149302
## figo        1.883627 6.503409
```

Yes, the results are the same

g. Calculate profile log likelihood confidence intervals.

```
# for log OR
confint(model.lr1)
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.2669574 0.7681153
## figo        0.6607034 1.9081579
```

```
# for OR
exp(confint(model.lr1))
```

```
## Waiting for profiling to be done...
```

```
##                2.5 %  97.5 %
## (Intercept) 1.305985 2.15570
## figo        1.936154 6.74066
```

There are slight differences. The confidence bounds are slightly higher for the profile method.

h. A logistic regression with diameter as independent variable.

```
model.lr2 <- glm(death~diameter, family=binomial, data=ova)
summary(model.lr2 )
```

```
##
## Call:
## glm(formula = death ~ diameter, family = binomial, data = ova)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.004807   0.186821  -0.026    0.979
## diameter     0.243340   0.048997   4.966 6.82e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 444.89  on 357  degrees of freedom
## Residual deviance: 417.69  on 356  degrees of freedom
## AIC: 421.69
##
## Number of Fisher Scoring iterations: 4
```

exp(B) = exp(0.243) = 1.276 . This indicates that for a centimeter increase in diameter of the tumor the odds of dying becomes 1.276 times higher. The p-value corresponding to the Wald test for H0: β =0 is very small indicating that there is a highly statistically significant association between tumor size and death.

i. the deviance (-2log-likelihood) of the model is 417.69, and the deviance of the null model is 444.89. The difference is 27.21. The corresponding p-value can be obtained from a chi-square distribution with 1 degree of freedom.

```
dif.deviance <- model.lr2$null.deviance-model.lr2$deviance
dif.deviance
```

```
## [1] 27.20842
```

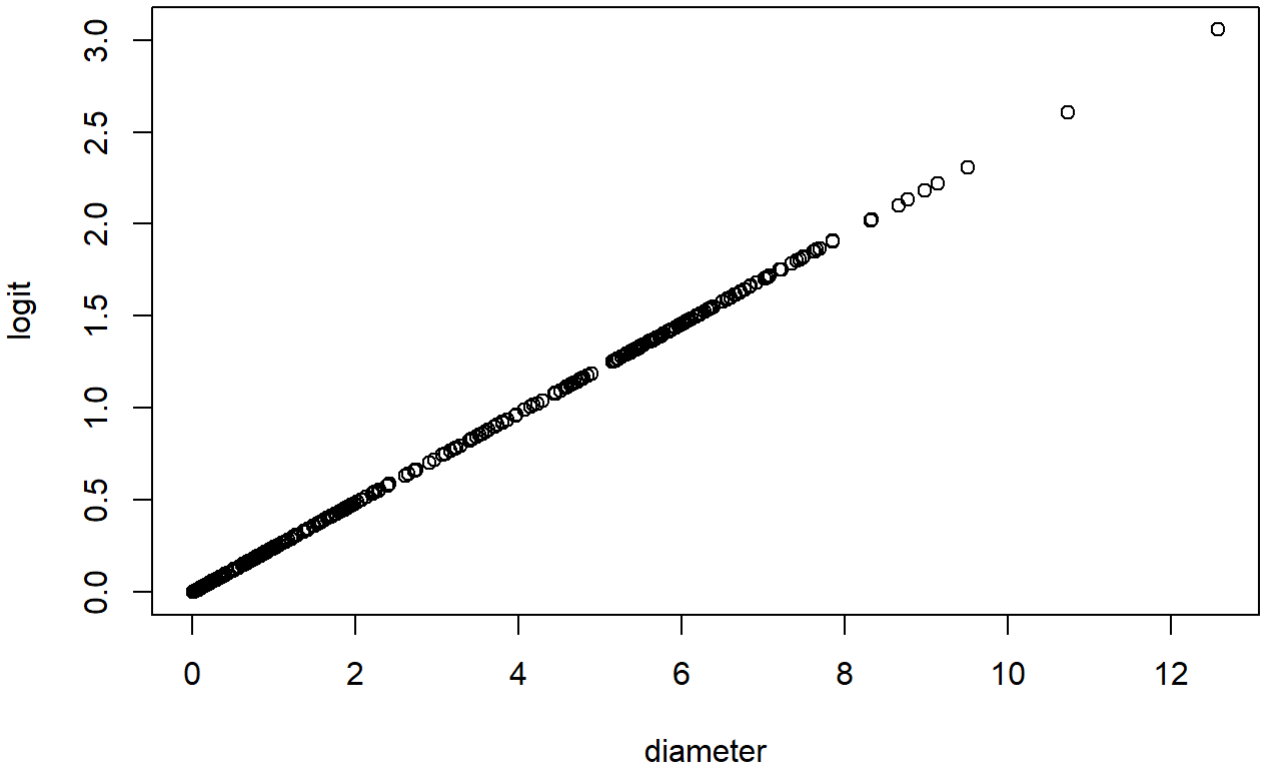
```
1-pchisq(dif.deviance, 1)
```

```
## [1] 1.826619e-07
```

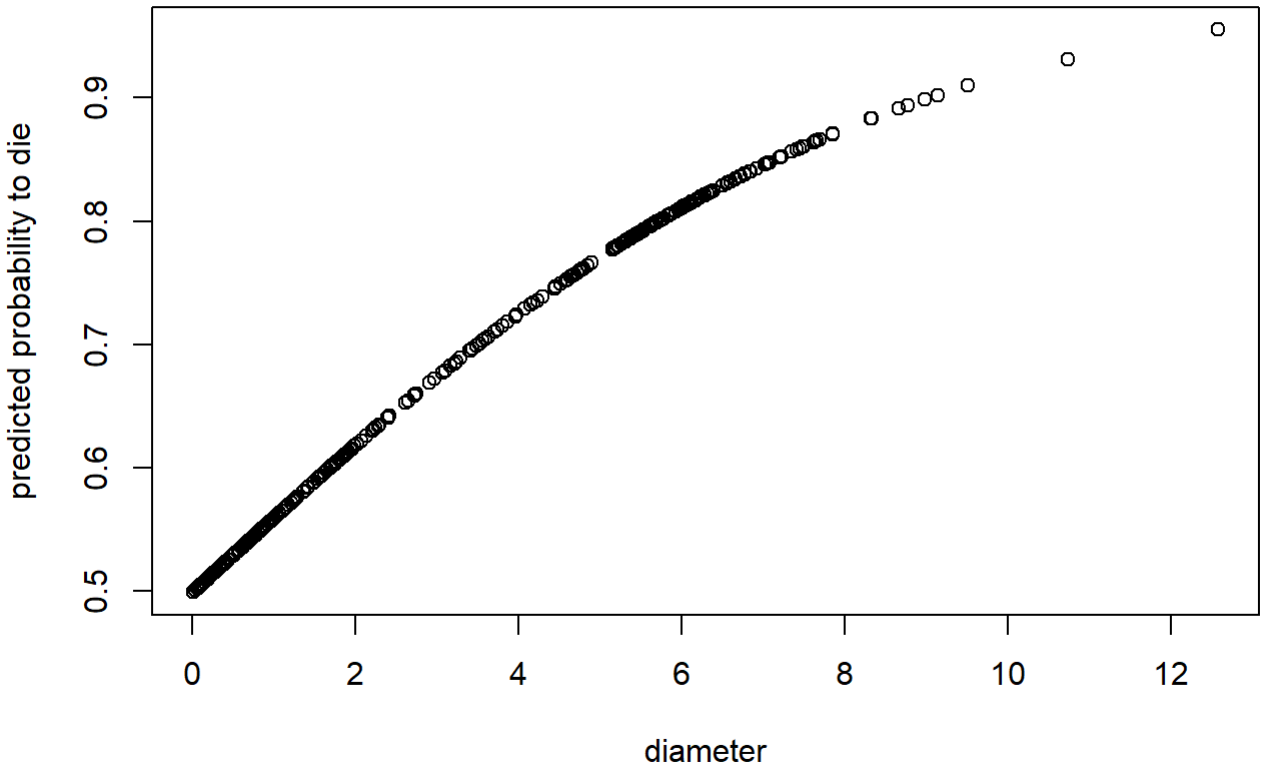
j. The logit for a tumor diameter of 5 cm is = -0.005 +0.243* 5 = 1.212. The probability to die within four years is P(death) = exp(1.212)/(1+exp(1.212)) = 0.771.

k. Calculate the predicted logit and predicted probability for each women.

```
# calculate the logit
ova$logit.model2 <- predict(model.lr2)
# calculate the predicted probability, by adding the option type="response"
ova$pred.model2 <- predict(model.lr2, type="response")
plot(ova$diameter, ova$logit.model2,xlab="diameter", ylab="logit")
```




```
plot(ova$diameter, ova$pred.model2,xlab="diameter", ylab="predicted probability to die")
```



l. A cross tabulation of death against ascites with estimated probability of death in the 3 categories.

```
tab <- xtabs(~ascites+death, data=ova)
tab
```

```
##        death
## ascites   0   1
##       0  14  38
##       1  42  52
##       2  56 156
```

```
proportions(tab, "ascites")
```

```
##        death
## ascites         0         1
##       0 0.2692308 0.7307692
##       1 0.4468085 0.5531915
##       2 0.2641509 0.7358491
```

m. Logistic regression with ascites as factor in the model.

```
model.lr3 <- glm(death~as.factor(ascites), family=binomial, data=ova)
summary(model.lr3 )
```

```
##
## Call:
## glm(formula = death ~ as.factor(ascites), family = binomial,
##     data = ova)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.99853    0.31264   3.194   0.0014 **
## as.factor(ascites)1 -0.78495    0.37521  -2.092   0.0364 *
## as.factor(ascites)2  0.02598    0.34930   0.074   0.9407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 444.89  on 357  degrees of freedom
## Residual deviance: 434.62  on 355  degrees of freedom
## AIC: 440.62
##
## Number of Fisher Scoring iterations: 4
```

Ascites=0 (unknown) is used as reference category. For someone with ascites = 0, the predicted probability to die within 4 years is exp(0.999) / (1+ exp(0.999)) = 0.731. For someone with ascites=1 (absent), the odds of dying is exp(-0.785)= 0.456 times lower. The probability to die in this group is exp(0.999 +-0.785) / (1+ exp(0.999 +-0.785)) = = 0.553. For someone with ascites = 2 (present), the odds of dying is exp(0.026)= 1.026 times higher. The probability to die in this group is exp(0.999 +0.026) / (1+ exp(0.999 +0.026)) = 0.736. The same probabilities as in the cross table.

n. A model with all covariates.

```
model.lr4 <- glm(death~as.factor(ascites)+ karn+figo+diameter, family=binomial, data=ova)
summary(model.lr4 )
```

```
##
## Call:
## glm(formula = death ~ as.factor(ascites) + karn + figo + diameter,
##     family = binomial, data = ova)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.65729    1.13073   1.466  0.14274
## as.factor(ascites)1 -0.30842    0.40567  -0.760  0.44709
## as.factor(ascites)2  0.25000    0.37522   0.666  0.50524
## karn                -0.01993    0.01164  -1.711  0.08702 .
## figo                 0.97861    0.33016   2.964  0.00304 **
## diameter             0.19810    0.05069   3.908 9.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 444.89  on 357  degrees of freedom
## Residual deviance: 395.72  on 352  degrees of freedom
## AIC: 407.72
##
## Number of Fisher Scoring iterations: 4
```

The coefficients for the two ascites categories have changed, and the regression coefficients for neither of the two categories differed significantly from the reference category. This is because the coefficients in a multi-variable model have a different interpretation. They are reflecting the effect of a variable, adjusted for the other variables.

Exercise 2

a. Make a 2x2 table of y against x.

```
library(readr)
data_logreg<- read_csv("data_logreg.csv", show_col_types = FALSE)

xtabs(~x+y, data=data_logreg)
```

```
##    y
## x     0   1
##   0 300   0
##   1  50 150
```

In the group X=0, there are no observations with Y=1, and P(Y=1|X=0)=0. The odds ratio of the two by two table is infinite.

b. Logistic regression

```
model.lr<-glm(y~x, data=data_logreg, family=binomial)
summary(model.lr)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial, data = data_logreg)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -20.57    1023.66  -0.020    0.984
## x              21.66    1023.66   0.021    0.983
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 610.86  on 499  degrees of freedom
## Residual deviance: 224.93  on 498  degrees of freedom
## AIC: 228.93
##
## Number of Fisher Scoring iterations: 19
```

We observe extremely large parameter estimates, very large standard errors and very large p-values. This indicates that there is something wrong. The model fitting algorithm did not converge.