Exercises to practice for exam Logistic regression, Poisson regression and generalized linear model

We consider data on mortality of new-born babies in their first year of life. The question is whether or not a mother's smoking during pregnancy puts a new-born baby at risk. The data are summarized in the Table below.

		Premature died in year 1	Premature alive at year 1	Full term died in year 1	Full term alive at year 1
Young mothers	Non-smokers	50	315	24	4012
	Smokers	9	40	6	459
Older mothers	Non-smokers	41	147	14	1594
	Smokers	4	11	1	124

The data were analyzed in R. The data set contains 6851 cases and the following variables:

AGE	=0 for young mothers, =1 for older mothers.
SMOKING	=0 for non-smokers, =1 for smokers.
PREM	=1 for premature, =0 for full term baby.
DEAD	=1 if baby died in 1st year, =0 if alive at year 1.

At the next page(s) you will find the output of three logistic regression models applied on this data.

a. Consider model1.

(i) Consider the covariate AGE. Is the effect of AGE significant? Give Wald's test statistic. Give also a 95% confidence interval for the oddsratio of AGE. You can use that $z_{0.025} = 1.96$.

The Wald statistic is b/se(b) = 0.4675/0.1803 = 2.59. That is larger than the critical value $z_{0.025} = 1.96$. Therefore we know that p-value < 0.05.

The odds ratio is exp(0.4675)=1.60. 95 % CI: exp(0.4675- 1.96*0.1803), exp(0.4675+ 1.96*0.1803)=(1.12 2.27)

(ii) Use the model to estimate of the probability that a premature new born baby dies in the first year of life, for an older mother who does not smoke.

Age=1, prem=1, smoking =0 → prob = exp(-5.1237 + 3.3098 + 0 + 0.4675)/(1+ exp(-5.1237 + 3.3098 + 0 + 0.4675) = 0.206

(iii) The given value of the maximized -2 Log Likelihood, can it be used as a goodness-of-fit test in this case?

In this case that would be possible because all covariates are categorical and there are several replicates with exactly the same covariate values (for example 50+315 observations in the category, young women, non smoking premature,

(iv) What is the conclusion from this model with respect to the influence of smoking? Is the conclusion warranted that smoking by the mother during pregnancy does not significantly increase the risk of dying of her baby in the first year of life?

The odds ratio is 1.53. We observe that for a given age group and prematurity, the odds on mortality is 1.53 times larger for smoking mothers. That effect is not statistically significant (p=0.10). This is a model which adjusts for premature and age.

b. Consider model 2. What is your conclusion with respect to the effect of smoking? Why is the result different? In order to answer the question whether or not a mother's smoking during pregnancy puts a new born baby at risk, what do you think that the more appropriate model is, model I or model II? Motivate your answer.

Here, the odds ratio is 1.56, slightly larger, with a smaller p-value (0.06). The p-value is larger than 0.05, (but only slightly). In this model there is no adjustment for prematurity. This may be more appropriate because smoking may cause prematurity, so you can discuss whether you would want to adjust for it (more on this in the course causal inference 1)

c. Consider model III. Almost all output of this model could be easily reproduced by hand. Fill in the numbers that should be at the places of the question marks.

For age = 0, smoking = 0 the risk on mortality is (50+24)/(50+315+24+4012)= 0.0168.
For age = 0, smoking = 1 the risk on mortality is (9+6)/(9+40+ 6 + 459)= 0.02918288.
For age = 1, smoking =0 the risk on mortality is 0.0306
For age = 1, smoking = 1 the risk on mortality is 0.0357

On the log odds scale
For age = 0, smoking = 0 the log odds on mortality is= log(0.0168/(1-0.0168))= -4.07 (= $\widehat{\beta_0}$)
For age = 0, smoking = 1 the log odds on mortality = log(0.02918288/(1-0.02918288))= -3.50 (= $\widehat{\beta_0} + \hat{\beta}_{smoking}$)
For age = 1, smoking =0 the log odds on mortality is -3.45 (= $\widehat{\beta_0} + \hat{\beta}_{age}$)

For age = 1, smoking = 1 the log odds on mortality is -3.30(= $\widehat{\beta_0} + \hat{\beta}_{smoking} + \hat{\beta}_{age} + \hat{\beta}_{interaction}$)

Therefore: the intercept is -4.07
The beta for smoking is -3.50 - -4.07= 0.57
The beta for age is -3.45 - -4.07= 0.62
The interaction term is -3.30-0.62-0.57- -4.07 =-0.42.

Let's check

```
Call:
glm(formula = death ~ age + smoking + age:smoking, family = binomial,
    data = infants)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0686     0.1172 -34.710  < 2e-16 ***
age           0.6137     0.1803   3.404 0.000663 ***
smoking       0.5640     0.2871   1.965 0.049449 *
age:smoking  -0.4050     0.5555  -0.729 0.465992
```

Very close, difference can be explained by rounding errors.

```
> model1<- glm(death~prem+smoking+age,family=binomial, data=infants)
> summary(model1)

Call:
glm(formula = death ~ prem + smoking + age, family = binomial,
    data = infants)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.1237     0.1690 -30.322  < 2e-16 ***
prem          3.3098     0.1846  17.929  < 2e-16 ***
smoking       0.4228     0.2624   1.611  0.10710
age           0.4675     0.1803   ……..  ……
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1435.5  on 6850  degrees of freedom
Residual deviance: 1084.8  on 6847  degrees of freedom
AIC: 1092.8

Number of Fisher Scoring iterations: 7

>
> model2 <- glm(death~smoking+age,family=binomial, data=infants)
> summary(model2)

Call:
glm(formula = death ~ smoking + age, family = binomial, data = infants)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0494     0.1132 -35.765  < 2e-16 ***
smoking       0.4438     0.2447   1.813 0.069764 .
age           0.5679     0.1697   3.347 0.000816 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1435.5  on 6850  degrees of freedom
Residual deviance: 1422.4  on 6848  degrees of freedom
AIC: 1428.4

Number of Fisher Scoring iterations: 6

>
> model3 <-glm(death~age+ smoking+ age:smoking,family=binomial, data=infant
s)
> summary(model3)
```

```
Call:
glm(formula = death ~ age + smoking + age:smoking, family = binomial,
    data = infants)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) ???
age         ???
smoking     ???
age:smoking ???
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1435.5  on 6850  degrees of freedom
Residual deviance: 1421.9  on 6847  degrees of freedom
AIC: 1429.9

Number of Fisher Scoring iterations: 6
```

Exercise 2 POISSON REGRESSION

We consider n = 287 ocean swimmers who were interviewed and data on the following variables were collected.

INF number of ear infections during last 5 years
SEX gender (0 = male, 1 = female)
LOC location of swimming (0 = beach, 1 = non-beach)
FOS frequent ocean swimmer (1 = yes, 0 = no)

One of the objectives of the study was whether beach swimmers run a greater risk of contracting ear infections than non-beach swimmers. On the following pages you will find some output of Poisson regressions carried out with R, with INF as the dependent variable, and one or more of the other variables as independent variables.

The data are organised at a person-by-person basis. As illustration the data of the first 10 individuals are listed.

INF	LOC	FOS	SEX
1	0	1	0
2	0	0	0
0	1	1	1
2	1	0	0
1	0	1	0
2	0	0	0
0	1	1	1
2	0	0	0
0	1	1	1
5	1	0	0
.	.	.	.
.	.	.	.

Some descriptive statistics are given in the following table.

SEX	FOS	LOC	number of infections	number of individuals
0	0	0	65	53
0	0	1	28	42
0	1	0	128	50
0	1	1	46	43
1	0	0	16	18
1	0	1	31	30
1	1	0	34	19
1	1	1	50	32

A Poisson model is fitted on these data. Study the output of model 1.

```
Call:
glm(formula = inf ~ loc + fos + sex, family = poisson, data = swimming)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.201746   0.097125   2.077   0.0378 *
loc         -0.509428   0.104176  -4.890 1.01e-06 ***
fos          0.612958   0.105001   5.838 5.29e-09 ***
sex          0.004485   0.108113   0.041   0.9669
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 824.51  on 286  degrees of freedom
Residual deviance: 764.65  on 283  degrees of freedom
AIC: 1145

Number of Fisher Scoring iterations: 6


>
```

1. What is the interpretation of the intercept?
   Is log of mean number of infections in the subgroup of male, beach, non frequent swimmers (SEX=0, FOS=0, LOC=0).

2. Show how you can calculate the intercept by yourself.

   There were 65 infections in 53 participants in the subgroup of male, beach, non frequent swimmers. Log (65/53) = 0.2040954, close to the intercept.

3. Explain why the number of degrees of freedom of the deviance of this model is 283.
   The saturated model has a parameter for each individual, so 287 parameters. The model at hand has 4 parameters, so the deviance has 287 – 4 = 283 df.

4. How can you see from the output that this is a badly fitting model?
   The deviance is rather large, compared to the number of degrees of freedom

5. Can you give an intuitive explanation why this is a badly fitting model?
   The model assumes that each person has the same infection rate, which is very unlikely.

Next two Poisson models with overdispersion is fitted.

b. Study the output of model 2A and 2B.

```
>
> model2a <- glm(inf~loc+fos,family=quasipoisson , data=swimming)
> summary(model2a)

Call:
glm(formula = inf ~ loc + fos, family = quasipoisson, data = swimming)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1266  -1.5652  -1.2137   0.5128   6.2538

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2029     0.1723   1.178  0.23981
loc          -0.5087     0.1903  -2.673  0.00795 **
fos           0.6130     0.1943   3.155  0.00178 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 3.426425)

    Null deviance: 824.51  on 286  degrees of freedom
Residual deviance: 764.65  on 284  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6

>
> model2b <- glm(inf~loc+fos+loc:fos,family=quasipoisson , data=swimming)
> summary(model2b)

Call:
glm(formula = inf ~ loc + fos + loc:fos, family = quasipoisson,
    data = swimming)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1669  -1.5105  -1.2802   0.5875   6.1649

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1318     0.2047   0.644   0.5203
loc          -0.3309     0.3153  -1.049   0.2949
fos           0.7217     0.2507   2.879   0.0043 **
loc:fos      -0.2757     0.3946  -0.699   0.4853
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 3.393966)

    Null deviance: 824.51  on 286  degrees of freedom
Residual deviance: 763.00  on 283  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

1. Explain what the dispersion parameter of 3.426425 means.

In this model the variance of Y is 3.42 times the mean of Y.

2. Is model 2A a well fitting model compared to b? Why?

The Wald test indicates that the interaction term is not statistically significant, so model 2A is a well fitting model compared to model 2B.

c. Finally a model with a different link function is fitted.

```
Call:
glm(formula = inf ~ loc + fos, family = poisson(link = "identity"),
    data = swimming)

Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
-2.0308  -1.6238  -1.2028   0.5511   6.4690

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.3184     0.1190  11.082  < 2e-16 ***
loc          -0.5950     0.1346  -4.420 9.89e-06 ***
fos           0.7436     0.1356   5.486 4.12e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 824.51  on 286  degrees of freedom
Residual deviance: 770.14  on 284  degrees of freedom
AIC: 1148.5

Number of Fisher Scoring iterations: 6
```

- What link function is used here?

An identity link

- What is in model 3 the interpretation of the coefficient of LOC? What is the conclusion of model 3 with regard to the swimming location?
- The mean number of infections is for non-beach swimmers -0.5950 lower than for swimmers. This is a statistically significantly difference.

3. In the journal Emerging Infectious Diseases, results from a study in 2020 regarding personal protective measures and risk for COVID-19 were studied. In this study 211 people who tested positive for Covid-19 were compared to 839 individuals, who never were tested positive. All participants were being asked whether they practiced social distancing, used face masks and how frequently they washed their hands in the past four weeks.

a. What kind of sampling design best describes this study? Prospective sampling (cohort study) versus retrospective sampling (case-control study).

retrospective sampling (case-control study)

Below are some results of the study.

Table 1

Handwashing	COVID-19 cases, no. (%), N = 210	Controls, no. (%), N = 826
None	44 (20.9)	121 (14.6)
Sometimes	114 (54.3)	333 (40.3)
Often	52 (24.8)	372 (45.0)

b. What effect measure would you prefer to study the relation between handwashing and COVID-19 (risk difference, relative risk or an odds ratio)? Why?

~~c.~~ A case-control study, so an odds ratio is appropriate

The authors also study the effect of wearing face masks on the risk for COVID-19, using logistic regression. The following odds ratios for COVID-19 versus not COVID-19 are reported:

Compliance with mask-wearing	Crude odds ratio (95% CI)*	Adjusted odds ratio (95% CI) **
Not wearing a mask	Referent	Referent
Wearing a mask sometimes	0.75 (0.37–1.52)	0.87 (0.41–1.84)
Always wearing a mask	0.16 (0.07–0.36)	0.23 (0.09–0.60)

* Logistic regression model with compliance with mask-wearing as independent variable and COVID-19 (yes/no) as dependent variable.

** From a logistic model with as independent variables: compliance with mask-wearing, gender, age group, contact place, shortest distance of contact, duration of contact at <1 m, sharing dishes or cups, sharing cigarettes, handwashing.

d. Explain why the adjusted odds ratios differ from the crude odds ratios. What could be the motivation of the researches to calculate adjusted odds ratios?

The researches may be interested in the effect of mask-wearing for a fixed value of gender, ager ….etc.

e. Are the effects of sometimes wearing a mask and always wearing statistically significantly different compared with not wearing a mask? Would you advice people to wear masks, based on these results?

Yes, there is a clear effect, monotone increasing. OR <1, significant for always wearing, which is associated with lower risk.