Exercises Day 4

PSY8003

Matthias Mittner

spring 2022

Exercise 1: Logistic regression

- as expected, survival probability drops for cheaper tickets and lower Ticket class
- when including `Pclass` the `Fare` effect is no longer significant (within each ticket class the tickets seem to have not varied much in price) but including both predictors is still better in terms of model fit
- `Pclass` and `Fare` add to the overall model-fit
- looking at the ORs, the drop is dramatic (94% reduced survival probability for 3rd vs. 1st class!); probably the cheap cabins were located deep within the ship?

```r
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.1 --

v ggplot2 3.3.5     v purrr   0.3.4
v tibble  3.1.6     v dplyr   1.0.7
v tidyr   1.1.4     v stringr 1.4.0
v readr   2.1.1     v forcats 0.5.1

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
library(lmtest)
```

```
Loading required package: zoo
```

```

Attaching package: 'zoo'
```

```
The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```r
titanic <- haven::read_dta("../data/titanic.dta") |> na.omit()

mod1 <- glm(Survived ~ Fare, data=titanic, family=binomial)
summary(mod1)
```

```

Call:
glm(formula = Survived ~ Fare, family = binomial, data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5870  -0.9073  -0.8714   1.3331   1.5741

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.896828   0.107616  -8.334  < 2e-16 ***
Fare         0.015997   0.002502   6.394 1.61e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 901.25  on 712  degrees of freedom
AIC: 905.25

Number of Fisher Scoring iterations: 5
```

```r
mod2 <- glm(Survived ~ factor(Pclass), data=titanic, family=binomial)
summary(mod2)
```

```
Call:
glm(formula = Survived ~ factor(Pclass), family = binomial, data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4607  -0.7399  -0.7399   0.9184   1.6908

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.6451     0.1543   4.180 2.92e-05 ***
factor(Pclass)2  -0.7261     0.2168  -3.350 0.000808 ***
factor(Pclass)3  -1.8009     0.1982  -9.086  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 869.81  on 711  degrees of freedom
AIC: 875.81

Number of Fisher Scoring iterations: 4
```

```r
mod3 <- glm(Survived ~ factor(Pclass) + Fare, data=titanic, family=binomial)
summary(mod3)
```

```
Call:
glm(formula = Survived ~ factor(Pclass) + Fare, family = binomial,
    data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0170  -0.7789  -0.7275   1.0612   1.7320

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.098422   0.252624   0.390   0.6968
factor(Pclass)2 -0.326010   0.259653  -1.256   0.2093
factor(Pclass)3 -1.345719   0.257121  -5.234 1.66e-07 ***
```

```
Fare              0.006827   0.002708   2.521   0.0117 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 861.50  on 710  degrees of freedom
AIC: 869.5

Number of Fisher Scoring iterations: 4
```

```
lrtest(mod1,mod2,mod3)
```

```
Likelihood ratio test

Model 1: Survived ~ Fare
Model 2: Survived ~ factor(Pclass)
Model 3: Survived ~ factor(Pclass) + Fare
  #Df  LogLik Df   Chisq Pr(>Chisq)
1   2 -450.63
2   3 -434.91  1 31.4428  2.054e-08 ***
3   4 -430.75  1  8.3077   0.003948 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod4 <- glm(Survived ~ Age*Sex + factor(Pclass) + Fare, data=titanic, family=binomial)
summary(mod4)
```

```
Call:
glm(formula = Survived ~ Age * Sex + factor(Pclass) + Fare, family = binomial,
    data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4957  -0.6416  -0.3714   0.6598   2.5934

Coefficients:
```

```
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      3.1432058  0.4960243   6.337 2.35e-10 ***
Age             -0.0072077  0.0116843  -0.617 0.537320    
Sexmale         -1.1585842  0.4385724  -2.642 0.008249 ** 
factor(Pclass)2 -1.5322436  0.3407047  -4.497 6.88e-06 ***
factor(Pclass)3 -2.7279214  0.3462059  -7.879 3.29e-15 ***
Fare            -0.0003952  0.0023346  -0.169 0.865561    
Age:Sexmale     -0.0501481  0.0148907  -3.368 0.000758 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 635.36  on 707  degrees of freedom
AIC: 649.36

Number of Fisher Scoring iterations: 5
```

```
exp(coef(mod4))
```

```
    (Intercept)             Age         Sexmale factor(Pclass)2 factor(Pclass)3 
     23.1780517       0.9928182       0.3139303       0.2160504       0.0653550 
           Fare     Age:Sexmale 
      0.9996048       0.9510885 
```

```
lrtest(mod3,mod4)
```

```
Likelihood ratio test

Model 1: Survived ~ factor(Pclass) + Fare
Model 2: Survived ~ Age * Sex + factor(Pclass) + Fare
  #Df  LogLik Df  Chisq Pr(>Chisq)    
1   4 -430.75                         
2   7 -317.68  3 226.15  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 2: Logistic regression for classification

- all variables except `make` are significant
- ORs are the multiplicative effect on the probability that an email is spam
- the `dollar` and `n000` variables have huge ORs
- the confusion matrix show perfect categorization (all emails are correctly categorized as spam or not)
- this is due to overfitting on the training dataset; a better way to test this is cross-validation (hold-out datasets)

```
spam <- haven::read_dta("../data/spam.dta")

mod <- glm(isspam ~ crltot + dollar + bang + money + n000 + make, data=spam, family=binomi
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(mod)
```

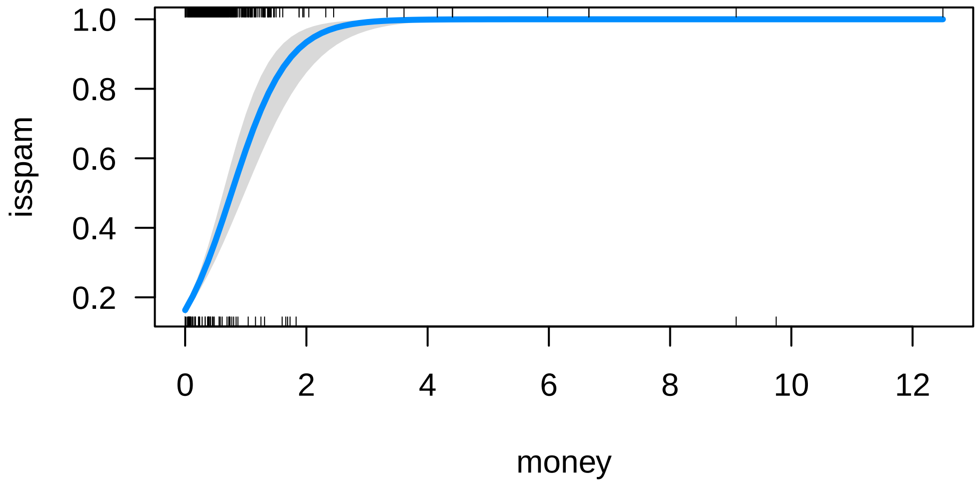
```
Call:
glm(formula = isspam ~ crltot + dollar + bang + money + n000 +
    make, family = binomial(), data = spam)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.6153  -0.5816   0.4439   1.9323

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.700e+00  5.361e-02 -31.717  < 2e-16 ***
crltot       6.917e-04  9.745e-05   7.098 1.27e-12 ***
dollar       8.013e+00  6.175e-01  12.976  < 2e-16 ***
bang         1.572e+00  1.115e-01  14.096  < 2e-16 ***
money        2.142e+00  2.418e-01   8.859  < 2e-16 ***
n000         4.149e+00  4.371e-01   9.492  < 2e-16 ***
make         1.698e-02  1.434e-01   0.118    0.906
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6170.2  on 4600  degrees of freedom
Residual deviance: 4058.8  on 4594  degrees of freedom
AIC: 4072.8

Number of Fisher Scoring iterations: 16
```

```
exp(coef(mod))
```

```
 (Intercept)       crltot       dollar         bang        money         n000 
   0.1826354    1.0006919 3018.4867001    4.8156296    8.5141268   63.3512064 
        make 
   1.0171234 
```

```
library(visreg)
library(modelr)
visreg(mod, scale = "response",xvar = "money" )
```



```
spam.pred <- spam |>
  add_predictions(mod, type = "response") |>
  mutate(pred.spam=case_when(pred>0.5 ~ 1,
                                   T ~ 0))
```

```
predicted <- factor(spam.pred$pred.spam,
                    labels=c("no spam", "spam"))
observed  <- factor(spam.pred$pred.spam,
                    labels=c("no spam", "spam"))
library(caret)
```

```
Loading required package: lattice


Attaching package: 'caret'


The following object is masked from 'package:purrr':

    lift
```

```
confusionMatrix(predicted, observed)
```

```
Confusion Matrix and Statistics

          Reference
Prediction no spam spam
   no spam    3273    0
   spam          0 1328

               Accuracy : 1
                 95% CI : (0.9992, 1)
    No Information Rate : 0.7114
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.7114
         Detection Rate : 0.7114
   Detection Prevalence : 0.7114
      Balanced Accuracy : 1.0000

       'Positive' Class : no spam
```

Exercise 3: Poisson regression

```
affairs <- haven::read_dta("../data/affairs.dta")

affairs |>
  mutate(female=as.integer(gender==2)) -> affairs

# a few example models...
mod <- glm(affairs ~ female*age, data=affairs, family = poisson)
summary(mod)
```

```

Call:
glm(formula = affairs ~ female * age, family = poisson, data = affairs)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.346  -1.758  -1.580  -1.394   6.186

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.489730   0.169745  -2.885  0.00391 **
female       0.281204   0.245582   1.145  0.25219
age          0.026358   0.004942   5.334 9.62e-08 ***
female:age  -0.008972   0.006845  -1.311  0.18993
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2925.5  on 600  degrees of freedom
Residual deviance: 2885.2  on 597  degrees of freedom
AIC: 3387.2
```

```
Number of Fisher Scoring iterations: 6
```

```r
mod <- glm(affairs ~ female*religiousness, data=affairs, family = poisson)
summary(mod)
```

```
Call:
glm(formula = affairs ~ female * religiousness, family = poisson,
    data = affairs)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.254  -1.936  -1.506  -1.227   6.630

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           1.23588    0.12612   9.799  < 2e-16 ***
female               -0.06668    0.17446  -0.382    0.702
religiousness        -0.30395    0.04291  -7.083 1.41e-12 ***
female:religiousness  0.04299    0.05881   0.731    0.465
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2925.5  on 600  degrees of freedom
Residual deviance: 2830.8  on 597  degrees of freedom
AIC: 3332.8

Number of Fisher Scoring iterations: 7
```

```r
mod <- glm(affairs ~ female*yearsmarried , data=affairs, family = poisson)
summary(mod)
```

```
Call:
glm(formula = affairs ~ female * yearsmarried, family = poisson,
    data = affairs)
```

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.130  -1.764  -1.360  -1.162   6.729

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.403473   0.105073  -3.840 0.000123 ***
female               0.091300   0.151603   0.602 0.547020
yearsmarried         0.081176   0.008938   9.082  < 2e-16 ***
female:yearsmarried -0.005737   0.012880  -0.445 0.656035
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2925.5  on 600  degrees of freedom
Residual deviance: 2766.4  on 597  degrees of freedom
AIC: 3268.4

Number of Fisher Scoring iterations: 7
```

```
exp(coef(mod))
```

```
        (Intercept)              female        yearsmarried female:yearsmarried
          0.6679960           1.0955976           1.0845621           0.9942798
```

```
library(AER)
```

```
Loading required package: car

Loading required package: carData


Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode
```

```
The following object is masked from 'package:purrr':

    some

Loading required package: sandwich

Loading required package: survival


Attaching package: 'survival'

The following object is masked from 'package:caret':

    cluster
```

```
dispersiontest(mod)
```

```

    Overdispersion test

data:  mod
z = 8.3884, p-value < 2.2e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  7.097836
```

```
library(MASS)
```

```

Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```
mod <- glm.nb(affairs ~ female*yearsmarried , data=affairs)
summary(mod)
```

```
Call:
glm.nb(formula = affairs ~ female * yearsmarried, data = affairs,
    init.theta = 0.1202267055, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8612  -0.7974  -0.7166  -0.6622   1.7407

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.455305   0.303679  -1.499  0.13380
female              -0.027146   0.449822  -0.060  0.95188
yearsmarried         0.086814   0.030411   2.855  0.00431 **
female:yearsmarried  0.006638   0.044534   0.149  0.88150
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.1202) family taken to be 1)

    Null deviance: 347.24  on 600  degrees of freedom
Residual deviance: 332.98  on 597  degrees of freedom
AIC: 1498.5

Number of Fisher Scoring iterations: 1

              Theta:  0.1202
          Std. Err.:  0.0129

 2 x log-likelihood:  -1488.5420
```

```
exp(coef(mod))
```

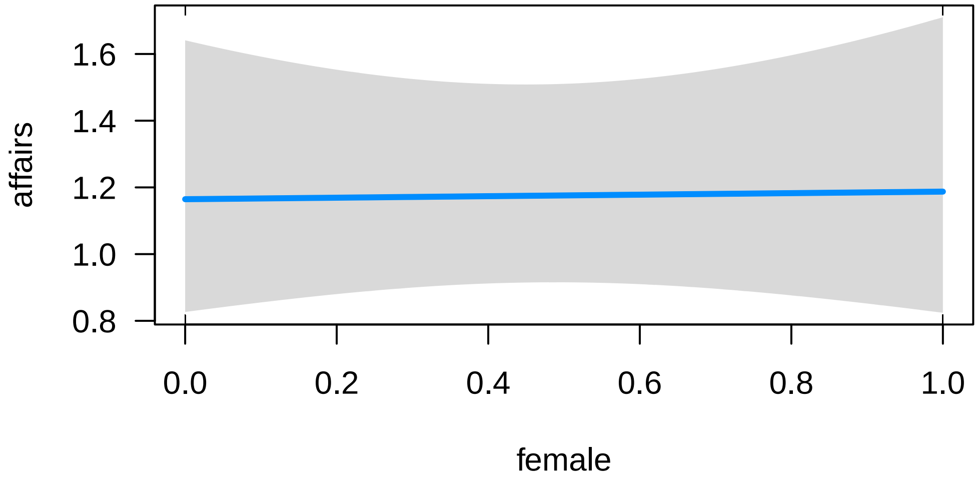
```
        (Intercept)              female        yearsmarried female:yearsmarried
          0.6342546           0.9732188           1.0906936           1.0066604
```

```
visreg(mod, scale="response")
```

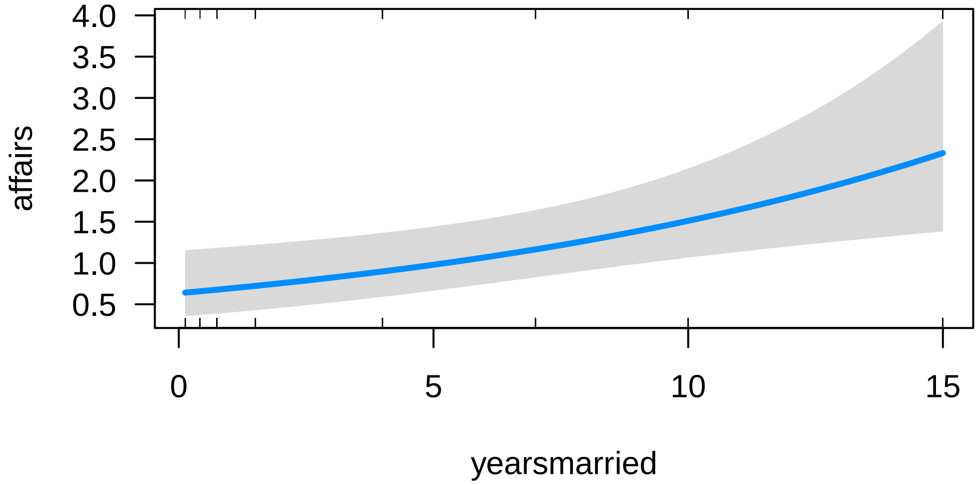
```
Warning:   Note that you are attempting to plot a 'main effect' in a model that contains an
  interaction.  This is potentially misleading; you may wish to consider using the 'by'
  argument.
```

```
Conditions used in construction of plot
yearsmarried: 7
```

```
Warning:   Note that you are attempting to plot a 'main effect' in a model that contains an
  interaction.  This is potentially misleading; you may wish to consider using the 'by'
  argument.
```



```
Conditions used in construction of plot
female: 0
```


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
affairs
0
5
10
15
yearsmarried