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?

```
use "../data/titanic.dta"
logistic Survived Fare, or
logistic Survived i.Pclass, or
logistic Survived Fare i.Pclass, or

gen Pclass2=0
replace Pclass2=1 if Pclass==2
gen Pclass3=0
replace Pclass3=1 if Pclass==3

logtest, m1(Survived Fare) m2(Survived Pclass2 Pclass3)
logtest, m1(Survived Fare) m2(Survived Fare Pclass2 Pclass3)

gen female=0
replace female=1 if Sex=="female"
summarize Age, meanonly
gen cAge = Age - r(mean)
generate cAge_female=cAge*female
```

```
logistic Survived female cAge cAge_female Fare Pclass2 Pclass3, or
```

```
Logistic regression                                     Number of obs =    891
                                                        LR chi2(1)    =  69.09
                                                        Prob > chi2   = 0.0000
Log likelihood = -558.78461                             Pseudo R2     = 0.0582

------------------------------------------------------------------------------
    Survived | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
        Fare |   1.015313   .0022662     6.81   0.000     1.010881    1.019764
       _cons |   .3901087   .0371125    -9.89   0.000     .3237485    .4700712
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.


Logistic regression                                     Number of obs =    891
                                                        LR chi2(2)    = 103.55
                                                        Prob > chi2   = 0.0000
Log likelihood = -541.55401                             Pseudo R2     = 0.0873

------------------------------------------------------------------------------
    Survived | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      Pclass |
          2  |   .5275925   .1076812    -3.13   0.002     .3536463    .7870968
          3  |    .188172   .0331014    -9.50   0.000     .1332968    .2656381
             |
       _cons |        1.7   .2395308     3.77   0.000     1.289776    2.240699
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.


Logistic regression                                     Number of obs =    891
                                                        LR chi2(3)    = 113.84
                                                        Prob > chi2   = 0.0000
Log likelihood = -536.40698                             Pseudo R2     = 0.0959

------------------------------------------------------------------------------
    Survived | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
```

```
        Fare |   1.006963   .0024621     2.84   0.005     1.002149      1.0118
             |
      Pclass |
          2  |   .7813293   .1873978    -1.03   0.304       .48829    1.250231
          3  |   .2920512   .0659187    -5.45   0.000     .1876441    .4545515
             |
       _cons |   .9944949   .2237613    -0.02   0.980      .639856    1.545692
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.


(277 real changes made)


(709 real changes made)


Likelihood ratio test for logit models

     m2 (unrestricted/extended model)
-----------------------------------------------------
    Survived | Coefficient  Std. err.      z    P>|z|
-------------+---------------------------------------
     Pclass2 |  -.6394311   .2040992    -3.13   0.002
     Pclass3 |  -1.670399   .1759104    -9.50   0.000
       _cons |   .5306283   .1409005     3.77   0.000
-----------------------------------------------------
Nobs:891     Pseudo-R2:0.09       LR chi2(2):103.55
                                     p-value:0.000

     m1 (restricted/parsimonous model)
-----------------------------------------------------
    Survived | Coefficient  Std. err.      z    P>|z|
-------------+---------------------------------------
        Fare |   .0151969    .002232     6.81   0.000
       _cons |  -.9413298   .0951337    -9.89   0.000
-----------------------------------------------------
Nobs:891     Pseudo-R2:0.06       LR chi2(1):69.09
                                     p-value:0.000

     m2 versus m1
-----------------------------------------------------
```

```
             LR-difference between m2 and m1: 34.46
                                  p-value: 0.000
-----------------------------------------------------


Likelihood ratio test for logit models

     m2 (unrestricted/extended model)
-----------------------------------------------------
    Survived | Coefficient  Std. err.      z    P>|z|
-------------+---------------------------------------
        Fare |   .0069391   .0024451     2.84   0.005
     Pclass2 |  -.2467586   .2398448    -1.03   0.304
     Pclass3 |  -1.230826   .2257092    -5.45   0.000
       _cons |  -.0055203   .2249999    -0.02   0.980
-----------------------------------------------------
Nobs:891     Pseudo-R2:0.10       LR chi2(3):113.84
                                     p-value:0.000

     m1 (restricted/parsimonous model)
-----------------------------------------------------
    Survived | Coefficient  Std. err.      z    P>|z|
-------------+---------------------------------------
        Fare |   .0151969    .002232     6.81   0.000
       _cons |  -.9413298   .0951337    -9.89   0.000
-----------------------------------------------------
Nobs:891     Pseudo-R2:0.06       LR chi2(1):69.09
                                     p-value:0.000

     m2 versus m1
-----------------------------------------------------

             LR-difference between m2 and m1: 44.76
                                  p-value: 0.000
-----------------------------------------------------


(466 real changes made)


(263 missing values generated)

(263 missing values generated)
```

```
Logistic regression                                     Number of obs =    714
                                                        LR chi2(6)    = 329.16
                                                        Prob > chi2   = 0.0000
Log likelihood = -317.67837                             Pseudo R2     = 0.3413

------------------------------------------------------------------------------
    Survived | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |   14.25442   3.131495    12.09   0.000     9.267289    21.92535
        cAge |    .944258   .0097088    -5.58   0.000     .9254197    .9634798
 cAge_female |   1.051427   .0156564     3.37   0.001     1.021184    1.082565
        Fare |   .9996048   .0023337    -0.17   0.866     .9950414    1.004189
     Pclass2 |   .2160504   .0736094    -4.50   0.000     .1108017     .421273
     Pclass3 |    .065355   .0226263    -7.88   0.000      .033158     .128816
       _cons |   1.310965   .3794314     0.94   0.350     .7434107    2.311818
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

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)

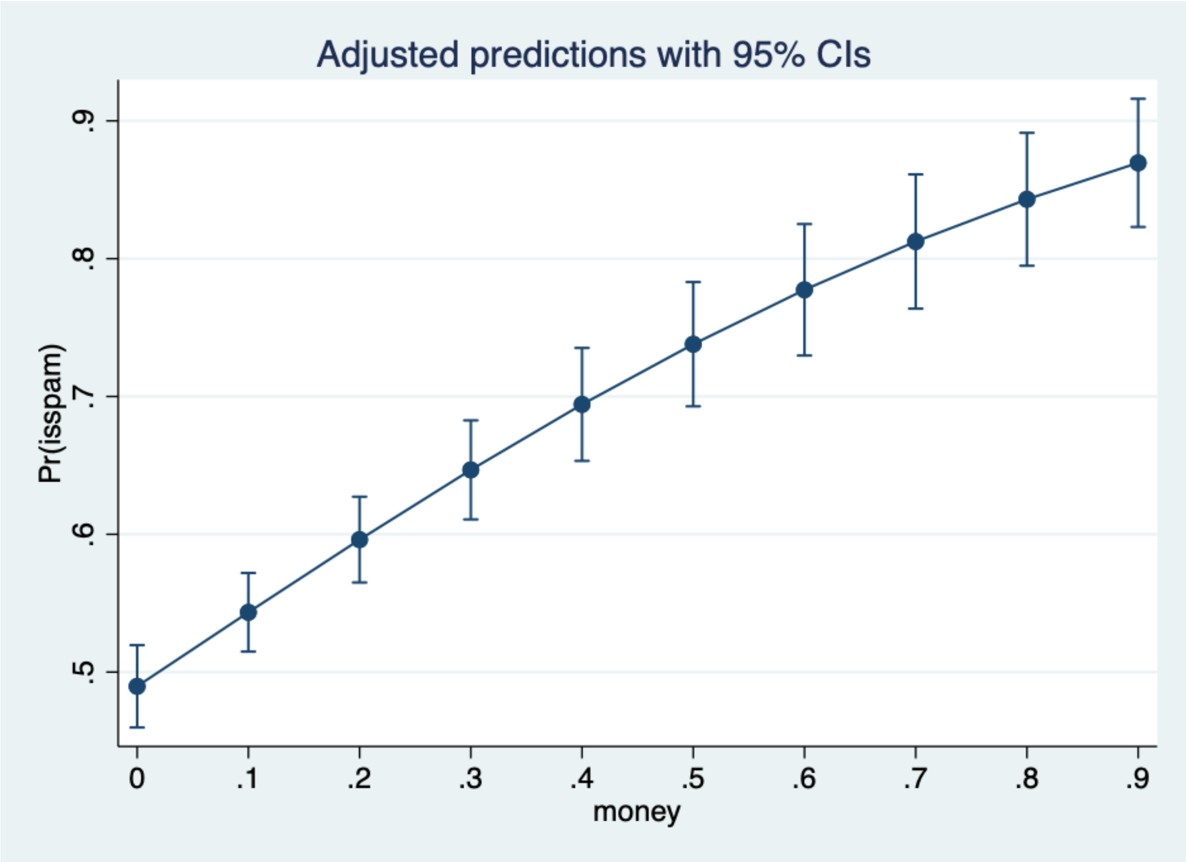
```
use "../data/spam.dta"
logistic isspam crltot dollar bang money n000 make, or iter(20)


margins,  atmeans at(money=(0(0.1)0.9))
marginsplot
quietly graph export pics/ex4_pred.png, replace

estat class
```

```
convergence not achieved

Logistic regression                                     Number of obs =   4,601
                                                        LR chi2(6)    = 2172.12
                                                        Prob > chi2   =  0.0000
Log likelihood = -1999.0182                             Pseudo R2     =  0.3520

------------------------------------------------------------------------------
      isspam | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      crltot |   1.000736   .0000991     7.42   0.000     1.000541     1.00093
      dollar |   4157.159    2621.16    13.22   0.000     1208.091    14305.19
        bang |   4.921111   .5521512    14.20   0.000     3.949643    6.131524
       money |   8.613476    2.09321     8.86   0.000     5.349621    13.86864
        n000 |   65.36898   28.92598     9.45   0.000     27.46075    155.6077
        make |   1.019318   .1474898     0.13   0.895     .7676185    1.353549
       _cons |   .1760463   .0095506   -32.02   0.000     .1582883    .1957967
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
Note: 0 failures and 28 successes completely determined.
Warning: Convergence not achieved.


Adjusted predictions                                      Number of obs = 4,601
Model VCE: OIM

Expression: Pr(isspam), predict()
1._at:  crltot = 283.2893 (mean)
        dollar = .0758107 (mean)
        bang   = .2690709 (mean)
        money  =        0
        n000   = .1016453 (mean)
        make   = .1045534 (mean)
2._at:  crltot = 283.2893 (mean)
        dollar = .0758107 (mean)
        bang   = .2690709 (mean)
        money  =       .1
        n000   = .1016453 (mean)
        make   = .1045534 (mean)
3._at:  crltot = 283.2893 (mean)
        dollar = .0758107 (mean)
        bang   = .2690709 (mean)
        money  =       .2
```

```
         n000   = .1016453 (mean)
         make   = .1045534 (mean)
4._at:   crltot = 283.2893 (mean)
         dollar = .0758107 (mean)
         bang   = .2690709 (mean)
         money  =       .3
         n000   = .1016453 (mean)
         make   = .1045534 (mean)
5._at:   crltot = 283.2893 (mean)
         dollar = .0758107 (mean)
         bang   = .2690709 (mean)
         money  =       .4
         n000   = .1016453 (mean)
         make   = .1045534 (mean)
6._at:   crltot = 283.2893 (mean)
         dollar = .0758107 (mean)
         bang   = .2690709 (mean)
         money  =       .5
         n000   = .1016453 (mean)
         make   = .1045534 (mean)
7._at:   crltot = 283.2893 (mean)
         dollar = .0758107 (mean)
         bang   = .2690709 (mean)
         money  =       .6
         n000   = .1016453 (mean)
         make   = .1045534 (mean)
8._at:   crltot = 283.2893 (mean)
         dollar = .0758107 (mean)
         bang   = .2690709 (mean)
         money  =       .7
         n000   = .1016453 (mean)
         make   = .1045534 (mean)
9._at:   crltot = 283.2893 (mean)
         dollar = .0758107 (mean)
         bang   = .2690709 (mean)
         money  =       .8
         n000   = .1016453 (mean)
         make   = .1045534 (mean)
10._at:  crltot = 283.2893 (mean)
         dollar = .0758107 (mean)
         bang   = .2690709 (mean)
         money  =       .9
         n000   = .1016453 (mean)
```

```
        make   = .1045534 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .4896631   .0152397    32.13   0.000     .4597938    .5195324
          2  |   .5433854   .0145584    37.32   0.000     .5148514    .5719194
          3  |   .5961171   .0158749    37.55   0.000     .5650028    .6272314
          4  |   .6467183   .0183257    35.29   0.000     .6108005     .682636
          5  |   .6942316   .0209072    33.21   0.000     .6532543    .7352089
          6  |   .7379441   .0230026    32.08   0.000     .6928599    .7830284
          7  |   .7774112   .0243358    31.95   0.000     .7297138    .8251085
          8  |    .812445   .0248468    32.70   0.000     .7637462    .8611437
          9  |   .8430778   .0246017    34.27   0.000     .7948593    .8912962
         10  |   .8695109    .023733    36.64   0.000     .8229952    .9160267
------------------------------------------------------------------------------


Variables that uniquely identify margins: money



Logistic model for isspam

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |      1182           152  |       1334
     -     |       631          2636  |       3267
-----------+--------------------------+-----------
   Total   |      1813          2788  |       4601

Classified + if predicted Pr(D) >= .5
True D defined as isspam != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   65.20%
Specificity                     Pr( -|~D)   94.55%
Positive predictive value       Pr( D| +)   88.61%
Negative predictive value       Pr(~D| -)   80.69%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    5.45%
```

```
False - rate for true D          Pr( -| D)   34.80%
False + rate for classified +    Pr(~D| +)   11.39%
False - rate for classified -    Pr( D| -)   19.31%
--------------------------------------------------
Correctly classified                         82.98%
--------------------------------------------------
```



Exercise 3: Poisson regression

```
use "../data/affairs.dta"
gen female=0
replace female=1 if gender==2
gen agefemale=female*age
gen relifemale=female*religiousness
```

```
poisson affairs female age agefemale, irr
poisson affairs female religiousness relifemale, irr

estat gof

overdisp affairs female religiousness relifemale
nbreg affairs female religiousness relifemale, irr
```

```
(286 real changes made)
```

```
Iteration 0:   log likelihood = -1689.5921
Iteration 1:   log likelihood =  -1689.592

Poisson regression                                      Number of obs =    601
                                                        LR chi2(3)    =  40.26
                                                        Prob > chi2   = 0.0000
Log likelihood = -1689.592                              Pseudo R2     = 0.0118

------------------------------------------------------------------------------
     affairs |        IRR   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |   1.324724   .3253462     1.14   0.252     .8186044    2.143762
         age |   1.026708    .005074     5.33   0.000     1.016812    1.036702
   agefemale |   .9910677   .0067842    -1.31   0.190     .9778598    1.004454
       _cons |   .6127915   .1040262    -2.88   0.004      .439354    .8546945
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.


Iteration 0:   log likelihood = -1662.3788
Iteration 1:   log likelihood = -1662.3788

Poisson regression                                      Number of obs =    601
                                                        LR chi2(3)    =  94.69
                                                        Prob > chi2   = 0.0000
Log likelihood = -1662.3788                             Pseudo R2     = 0.0277

-------------------------------------------------------------------------------
```

```
      affairs |        IRR   Std. err.      z    P>|z|     [95% conf. interval]
--------------+----------------------------------------------------------------
       female |   .9354976   .1632069    -0.38   0.702     .6645721    1.316871
religiousness |   .7379008   .0316627    -7.08   0.000     .6783811    .8026428
   relifemale |   1.043923   .0613956     0.73   0.465     .9302661    1.171466
        _cons |    3.44139   .4340443     9.80   0.000     2.687672    4.406476
-------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.

         Deviance goodness-of-fit =  2830.768
         Prob > chi2(597)         =    0.0000

         Pearson goodness-of-fit  =  4411.892
         Prob > chi2(597)         =    0.0000


Overdispersion test (H0: equidispersion)                   Number of obs = 601
------------------------------------------------------------------------------
     affairs | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        uhat |   3.897155   .4752853     8.20   0.000      2.96373     4.83058
------------------------------------------------------------------------------


Fitting Poisson model:

Iteration 0:   log likelihood = -1662.3788
Iteration 1:   log likelihood = -1662.3788

Fitting constant-only model:

Iteration 0:   log likelihood = -997.50487
Iteration 1:   log likelihood = -796.92568
Iteration 2:   log likelihood = -758.30801
Iteration 3:   log likelihood = -751.19633
Iteration 4:   log likelihood = -751.17313
Iteration 5:   log likelihood = -751.17313

Fitting full model:

Iteration 0:   log likelihood = -747.79297
Iteration 1:   log likelihood = -747.60846
Iteration 2:   log likelihood =  -747.6076
```

```
Iteration 3:   log likelihood =  -747.6076

Negative binomial regression                            Number of obs =    601
                                                        LR chi2(3)    =   7.13
Dispersion: mean                                        Prob > chi2   = 0.0678
Log likelihood = -747.6076                              Pseudo R2     = 0.0047

------------------------------------------------------------------------------
     affairs |        IRR   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |    1.08758    .780331     0.12   0.907     .2665201    4.438052
religiousness |   .7503675   .1124546    -1.92   0.055     .5593811    1.006561
   relifemale |   .9912554   .2159717    -0.04   0.968     .6467381    1.519297
       _cons |   3.278795   1.614518     2.41   0.016     1.249028    8.607087
-------------+----------------------------------------------------------------
    /lnalpha |   2.152435   .1067911                      1.943128    2.361742
-------------+----------------------------------------------------------------
       alpha |   8.605786   .9190218                      6.980552    10.60941
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to incidence-rate ratios.
Note: _cons estimates baseline incidence rate.
LR test of alpha=0: chibar2(01) = 1829.54              Prob >= chibar2 = 0.000
```