Problem Set 6, Winter 2021

Adam Ten Hoeve

```
knitr::opts_chunk$set(echo = TRUE)

# Load any packages, if any, that you use as part of your answers here
# For example:
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

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

CONTEXT: Pew Research Center data

The data in "pew_data.RData" comes from the Pew Research Center, an organization that conducts nationally-representative public opinion polls on a variety of political and social topics. Dr. Durso constructed this data set from the 2017 Pew Research Center Science and NewsSurvey, downloaded from https://www.journalism.org/datasets/2018/ on 4/16/2019.

The variable "LIFE" contains the responses of participants to the following question:

"In general, would you say life in America today is better, worse or about the same as it was 50 years ago for people like you?"

1 = Better today

2 = Worse today

3 = About the same as it was 50 years ago

-1 = Refused

You will use the pew data set again for these questions, but the set of variables will be different than those used in Problem Set 5. The data for these questions is in a data set called "pew2". Please run the code chunk below before starting this problem set (you will need the tidyverse package loaded into memory before running this chunk)

```
# Your working directory will need to be set to where the pew_data.RData is located on your computer

load("pew_data.RData")
pew2<-dplyr::select(dat,AGE,PPREG4,PPWORK,PPINCIMP,PPGENDER,PPETHM,IDEO,PPEDUCAT,LIFE, KNOWLEDGE,ENJOY,S
```

Question 1 - 5 points

Two of the new variables relate to use of social media. SNSUSE asks if the participant uses social media, and SNSFREQ asks how frequently the participant uses social media. Many of the NAs in this data set come from people who responded that they did not use social media; that is, these responses are not really missing.

To fix this, recode all NAs in SNSFREQ to 6 if the participant responded "no" to the SNSUSE variable. After doing this, please display the counts for responses to SNSFREQ and SNSUSE using the table() function (hint: you should be able to confirm that your recoding was done correctly using the information you get from the table() function).

```
# Your recoding code here
pew2$SNSUSE = as.factor(pew2$SNSUSE)
pew2$SNSFREQ = as.numeric(pew2$SNSFREQ)

pew2[(is.na(pew2$SNSFREQ)) & (pew2$SNSUSE==2), ]$SNSFREQ <- 6

# Don't forget to display the counts for SNSFREQ (with recoded value included) and SNSUSE
table(pew2$SNSUSE)
```

```
##
##   -1    1    2
##   12 2755 1257
```

```
table(pew2$SNSFREQ)
```

```
##
##   -1    1    2    3    4    5    6
##    6 1425  650  420  137  117 1257
```

Question 2 - 10 points

For this analysis, you will conduct a "complete case" analysis. That is, there will be no missing data in your data set at the start of your analysis. Be sure that you have completed Question 1 before starting this question, and then do the following steps in order:

1) Examine your variables to see what responses correspond to missing values. The attributes() and table() functions are useful for this, and examples of their use are shown in Problem Set 5. Consider labels such as "Not asked" and "Refused" as missing.

```
# Your code for variable examination here

table(pew2$PPINCIMP)
```

```
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##  66  31  40  91  77 104 144 183 179 167 258 321 378 285 319 486 226 253 160 125
##  21
## 131
```

```r
table(pew2$PPGENDER)
```

```
##
##    1    2
## 1993 2031
```

```r
table(pew2$PPETHM)
```

```
##
##    1    2    3    4    5
## 2862  392  166  447  157
```

```r
table(pew2$IDEO)
```

```
##
##   -1    1    2    3    4    5
##  116  314 1095 1624  616  259
```

```r
table(pew2$PPEDUCAT)
```

```
##
##    1    2    3    4
##  303 1130 1147 1444
```

```r
table(pew2$LIFE)
```

```
##
##   -1    1    2    3
##   18 1596 1900  510
```

```r
table(pew2$AGE)
```

```
##
##  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37
##  43  26  38  29  38  36  38  59  62  78  83  80  63  39  54  60  46  64  77  60
##  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57
##  60  57  74  55  51  55  66  45  80  72  69  60  72  71  68  82 101 109  92  93
##  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77
## 100 127  80  81  90  73  74  64  75  78  64  71  67  53  56  62  50  35  31  25
##  78  79  80  81  82  83  84  85  86  87  88  89  90
##  27  24  27  23  12  14   6  13   7   1   3   3   3
```

```r
table(pew2$PPREG4)
```

```
##
##    1    2    3    4
##  738  879 1485  922
```

```r
table(pew2$PPWORK)
```

```
##
##    1    2    3    4    5    6    7
## 2204  333   13  198  841  157  278
```

```r
table(pew2$KNOWLEDGE)
```

```
##
##   -1    1    2    3    4
##   13  411 2236 1174  190
```

```r
table(pew2$ENJOY)
```

```
##
##   -1    1    2    3    4
##   46  325 1775 1379  499
```

```r
table(pew2$SNSUSE)
```

```
##
##   -1    1    2
##   12 2755 1257
```

```r
table(pew2$SNSFREQ)
```

```
##
##   -1    1    2    3    4    5    6
##    6 1425  650  420  137  117 1257
```

```r
attributes(pew2$PPINCIMP)$labels
```

```
##            Not asked              REFUSED     Less than $5,000
##                   -2                   -1                    1
##     $5,000 to $7,499     $7,500 to $9,999   $10,000 to $12,499
##                    2                    3                    4
##   $12,500 to $14,999   $15,000 to $19,999   $20,000 to $24,999
##                    5                    6                    7
##   $25,000 to $29,999   $30,000 to $34,999   $35,000 to $39,999
##                    8                    9                   10
##   $40,000 to $49,999   $50,000 to $59,999   $60,000 to $74,999
##                   11                   12                   13
##   $75,000 to $84,999   $85,000 to $99,999 $100,000 to $124,999
##                   14                   15                   16
## $125,000 to $149,999 $150,000 to $174,999 $175,000 to $199,999
##                   17                   18                   19
## $200,000 to $249,999     $250,000 or more
##                   20                   21
```

```r
attributes(pew2$PPGENDER)$labels
```

```
## Not asked   REFUSED      Male    Female
##        -2        -1         1         2
```

```r
attributes(pew2$PPETHM)$labels
```

```
##              Not asked                REFUSED    White, Non-Hispanic
##                     -2                     -1                       1
##    Black, Non-Hispanic    Other, Non-Hispanic               Hispanic
##                      2                      3                       4
## 2+ Races, Non-Hispanic
##                      5
```

```r
attributes(pew2$IDEO)$labels
```

```
##           Refused Very conservative      Conservative          Moderate
##                -1                 1                 2                 3
##           Liberal      Very liberal
##                 4                 5
```

```r
attributes(pew2$PPEDUCAT)$labels
```

```
##                   Not asked                     REFUSED
##                          -2                          -1
##       Less than high school                 High school
##                           1                           2
##               Some college Bachelor's degree or higher
##                           3                           4
```

```r
attributes(pew2$LIFE)$labels
```

```
##                              Refused                           Better today
##                                   -1                                      1
##                          Worse today About the same as it was 50 years ago
##                                    2                                      3
```

```r
attributes(pew2$AGE)$labels
```

```
## 90 years or older
##                90
```

```r
attributes(pew2$PPREG4)$labels
```

```
## Not asked   REFUSED Northeast   Midwest     South      West
##        -2        -1         1         2         3         4
```

```
attributes(pew2$PPWORK)$labels
```

```
##                                     Not asked
##                                            -2
##                                      REFUSED
##                                            -1
##                  Working - as a paid employee
##                                             1
##                       Working - self-employed
##                                             2
## Not working - on temporary layoff from a job
##                                             3
##                Not working - looking for work
##                                             4
##                          Not working - retired
##                                             5
##                        Not working - disabled
##                                             6
##                           Not working - other
##                                             7
```

```
attributes(pew2$KNOWLEDGE)$labels
```

```
##        Refused          A lot           Some       Not much Nothing at all
##             -1              1              2              3              4
```

```
attributes(pew2$ENJOY)$labels
```

```
##                     Refused A lot more than other news
##                          -1                           1
##       More than other news       Less than other news
##                           2                           3
## A lot less than other news
##                           4
```

2) Count the number of observations (i.e., rows) in your data set.

```
# Your code here
nrow(pew2)
```

```
## [1] 4024
```

Number of rows in your data set (your answer here): 4024

3) Set these responses equal to "NA", which is R's internal marker for missing data.

```
# Your code here
# Set any value of -1 or -2 to NA
pew2 = pew2 %>% replace(.==-1 | .==-2, NA)
```

4) Remove all observations with NA responses from your data.

```
# Your code here
pew2 = drop_na(pew2)
```

5) Count the number of observations again.

```
# Your code here
nrow(pew2)
```

```
## [1] 3836
```

Number of rows in your complete-cases data set (your answer here): 3836

Question 3 - 5 points

Be sure that you have completed Question 2 before starting this question.

1) Recode the LIFE variable such that "Worse today" equals 1 and the other responses are equal to zero.
2) Change the variables to the appropriate variable type:
    - Continuous: age, PPINCIMP
    - Categorical: all others

```
# Your code here
pew2$LIFE = ifelse(pew2$LIFE==2, 1, 0)

pew2$LIFE = as.factor(pew2$LIFE)
pew2$PPGENDER = as.factor(pew2$PPGENDER)
pew2$PPETHM = as.factor(pew2$PPETHM)
pew2$PPEDUCAT = as.factor(pew2$PPEDUCAT)
pew2$PPREG4 = as.factor(pew2$PPREG4)
pew2$PPWORK = as.factor(pew2$PPWORK)
pew2$IDEO = as.factor(pew2$IDEO)
pew2$KNOWLEDGE = as.factor(pew2$KNOWLEDGE)
pew2$ENJOY = as.factor(pew2$ENJOY)
pew2$SNSUSE = as.factor(pew2$SNSUSE)
pew2$SNSFREQ = as.factor(pew2$SNSFREQ)

pew2$PPINCIMP = as.numeric(pew2$PPINCIMP)
pew2$AGE = as.numeric(pew2$AGE)

str(pew2)
```

```
## tibble [3,836 x 13] (S3: tbl_df/tbl/data.frame)
##  $ AGE      : num [1:3836] 64 32 58 46 34 23 26 29 68 26 ...
##  $ PPREG4   : Factor w/ 4 levels "1","2","3","4": 4 1 3 3 4 3 1 3 1 1 ...
##  $ PPWORK   : Factor w/ 7 levels "1","2","3","4",..: 1 1 5 1 2 4 1 2 5 1 ...
##  $ PPINCIMP : num [1:3836] 16 19 12 12 21 18 19 16 7 10 ...
##  $ PPGENDER : Factor w/ 2 levels "1","2": 1 2 1 1 1 1 2 2 2 2 ...
##  $ PPETHM   : Factor w/ 5 levels "1","2","3","4",..: 1 2 4 4 1 5 1 5 1 1 ...
##  $ IDEO     : Factor w/ 5 levels "1","2","3","4",..: 1 3 2 3 2 3 2 3 3 2 ...
##  $ PPEDUCAT : Factor w/ 4 levels "1","2","3","4": 4 4 2 1 3 3 4 4 2 4 ...
```

```
##  $ LIFE     : Factor w/ 2 levels "0","1": 2 2 1 2 2 1 2 1 2 2 ...
##  $ KNOWLEDGE: Factor w/ 4 levels "1","2","3","4": 2 3 2 3 1 3 2 2 3 1 ...
##  $ ENJOY    : Factor w/ 4 levels "1","2","3","4": 2 4 1 3 1 2 3 1 3 1 ...
##  $ SNSUSE   : Factor w/ 3 levels "-1","1","2": 2 3 2 3 2 2 2 2 3 2 ...
##  $ SNSFREQ  : Factor w/ 6 levels "1","2","3","4",..: 1 6 2 6 1 1 2 1 6 1 ...
```

Question 4 - 5 points

Split your data set into training, validation, and test sets. Use the following proportions: 70% training, 15% validation, and 15% test

```
# Your code here
train_size = floor(nrow(pew2) * 0.70)
valid_size = floor(nrow(pew2) * 0.15)

pew2.train = pew2[1:train_size, ]
pew2.valid = pew2[train_size:(train_size+valid_size), ]
pew2.test = pew2[(train_size+valid_size): nrow(pew2), ]
```

Question 5 - 5 points

Develop a set of candidate models by using forward selection to fit logistic regression using the binarization of LIFE as the outcome and all other variables in the data set as potential predictors. Display each step of the forward selection using the TRACE option.

```
# Code for your forward selection here

full.model.formula = as.formula("LIFE~AGE+PPREG4+PPWORK+PPINCIMP+PPGENDER+PPETHM+IDEO+PPEDUCAT+KNOWLEDG

pew.model.forward = step(glm(LIFE~1, pew2.train, family="binomial"),
                         scope=full.model.formula,
                         direction="forward",
                         trace=1)
```

```
## Start:  AIC=3713.56
## LIFE ~ 1
##
##             Df Deviance    AIC
## + PPEDUCAT   3   3649.6 3657.6
## + PPINCIMP   1   3662.6 3666.6
## + IDEO       4   3687.0 3697.0
## + PPGENDER   1   3701.1 3705.1
## + KNOWLEDGE  3   3699.0 3707.0
## + ENJOY      3   3700.0 3708.0
## + SNSUSE     1   3709.4 3713.4
## + PPREG4     3   3705.5 3713.5
## <none>           3711.6 3713.6
## + PPETHM     4   3704.5 3714.5
## + PPWORK     6   3700.6 3714.6
## + AGE        1   3711.4 3715.4
## + SNSFREQ    5   3703.8 3715.8
##
```

```
## Step:  AIC=3657.57
## LIFE ~ PPEDUCAT
##
##             Df Deviance    AIC
## + PPINCIMP   1   3627.3 3637.3
## + IDEO       4   3630.5 3646.5
## + PPGENDER   1   3641.8 3651.8
## + PPETHM     4   3636.8 3652.8
## + PPREG4     3   3642.9 3656.9
## <none>           3649.6 3657.6
## + ENJOY      3   3643.7 3657.7
## + SNSUSE     1   3648.0 3658.0
## + AGE        1   3649.5 3659.5
## + KNOWLEDGE  3   3646.4 3660.4
## + SNSFREQ    5   3643.6 3661.6
## + PPWORK     6   3641.9 3661.9
##
## Step:  AIC=3637.31
## LIFE ~ PPEDUCAT + PPINCIMP
##
##             Df Deviance    AIC
## + IDEO       4   3605.7 3623.7
## + PPETHM     4   3609.7 3627.7
## + PPGENDER   1   3621.4 3633.4
## + PPREG4     3   3618.4 3634.4
## <none>           3627.3 3637.3
## + SNSUSE     1   3626.1 3638.1
## + ENJOY      3   3622.4 3638.4
## + AGE        1   3627.3 3639.3
## + KNOWLEDGE  3   3625.0 3641.0
## + SNSFREQ    5   3621.6 3641.6
## + PPWORK     6   3620.4 3642.4
##
## Step:  AIC=3623.7
## LIFE ~ PPEDUCAT + PPINCIMP + IDEO
##
##             Df Deviance    AIC
## + PPGENDER   1   3596.9 3616.9
## + PPETHM     4   3591.4 3617.4
## + PPREG4     3   3595.5 3619.5
## <none>           3605.7 3623.7
## + SNSUSE     1   3604.2 3624.2
## + ENJOY      3   3601.3 3625.3
## + AGE        1   3605.6 3625.6
## + KNOWLEDGE  3   3602.8 3626.8
## + SNSFREQ    5   3599.3 3627.3
## + PPWORK     6   3597.4 3627.4
##
## Step:  AIC=3616.93
## LIFE ~ PPEDUCAT + PPINCIMP + IDEO + PPGENDER
##
##             Df Deviance    AIC
## + PPETHM     4   3583.0 3611.0
## + PPREG4     3   3585.7 3611.7
```

```
## <none>           3596.9 3616.9
## + PPWORK     6   3586.1 3618.1
## + SNSUSE     1   3596.3 3618.3
## + AGE        1   3596.9 3618.9
## + ENJOY      3   3593.5 3619.5
## + KNOWLEDGE  3   3595.2 3621.2
## + SNSFREQ    5   3592.0 3622.0
##
## Step:  AIC=3610.97
## LIFE ~ PPEDUCAT + PPINCIMP + IDEO + PPGENDER + PPETHM
##
##             Df Deviance    AIC
## + PPREG4     3   3573.3 3607.3
## <none>           3583.0 3611.0
## + PPWORK     6   3571.5 3611.5
## + SNSUSE     1   3582.5 3612.5
## + AGE        1   3582.7 3612.7
## + ENJOY      3   3579.1 3613.1
## + KNOWLEDGE  3   3581.1 3615.1
## + SNSFREQ    5   3577.4 3615.4
##
## Step:  AIC=3607.34
## LIFE ~ PPEDUCAT + PPINCIMP + IDEO + PPGENDER + PPETHM + PPREG4
##
##             Df Deviance    AIC
## <none>           3573.3 3607.3
## + PPWORK     6   3561.5 3607.5
## + SNSUSE     1   3572.8 3608.8
## + AGE        1   3572.9 3608.9
## + ENJOY      3   3569.7 3609.7
## + KNOWLEDGE  3   3571.5 3611.5
## + SNSFREQ    5   3568.0 3612.0
```

Question 6 - 10 points

Apply each of the models in your forward regression (as shown by the TRACE option) to the validation set. Compute the deviances of these models (hint: there is a good example of this in the async material in 5.2.1: backward_train_validate_test_5_2_1). Be sure to display the deviances for each model. Once you have the deviances, choose the best of these models.

```r
# Create the models for each step of the forward selection
pew.model.0 = glm(LIFE~1, pew2.train, family="binomial")
pew.model.1 = glm(LIFE~PPEDUCAT, pew2.train, family="binomial")
pew.model.2 = glm(LIFE~PPEDUCAT+PPINCIMP, pew2.train, family="binomial")
pew.model.3 = glm(LIFE~PPEDUCAT+PPINCIMP+IDEO, pew2.train, family="binomial")
pew.model.4 = glm(LIFE~PPEDUCAT+PPINCIMP+IDEO+PPGENDER, pew2.train, family="binomial")
pew.model.5 = glm(LIFE~PPEDUCAT+PPINCIMP+IDEO+PPGENDER+PPETHM, pew2.train, family="binomial")
pew.model.6 = glm(LIFE~PPEDUCAT+PPINCIMP+IDEO+PPGENDER+PPETHM+PPREG4, pew2.train, family="binomial")

models = c(pew.model.0, pew.model.1, pew.model.2, pew.model.3, pew.model.4, pew.model.5, pew.model.6)

# From 5.2.1
valid.dev <- function(m.pred, dat.this){
```

```
      pred.m <- predict(m.pred, newdata=dat.this, type="response")
      return(-2*sum(dat.this$LIFE*log(pred.m)+(1-dat.this$LIFE)*log(1-pred.m)))
}

# Convert LIFE to an int for the deviance equation to work
pew2.valid$LIFE = as.integer(pew2.valid$LIFE)-1

# Calculate the deviances for each model on the validation set
# Display the results
m.0.dev = valid.dev(pew.model.0, pew2.valid)
print(pew.model.0$formula)
```

```
## LIFE ~ 1
```

```
print(m.0.dev)
```

```
## [1] 797.0105
```

```
m.1.dev = valid.dev(pew.model.1, pew2.valid)
print(pew.model.1$formula)
```

```
## LIFE ~ PPEDUCAT
```

```
print(m.1.dev)
```

```
## [1] 783.0119
```

```
m.2.dev = valid.dev(pew.model.2, pew2.valid)
print(pew.model.2$formula)
```

```
## LIFE ~ PPEDUCAT + PPINCIMP
```

```
print(m.2.dev)
```

```
## [1] 777.0453
```

```
m.3.dev = valid.dev(pew.model.3, pew2.valid)
print(pew.model.3$formula)
```

```
## LIFE ~ PPEDUCAT + PPINCIMP + IDEO
```

```
print(m.3.dev)
```

```
## [1] 779.7715
```

```r
m.4.dev = valid.dev(pew.model.4, pew2.valid)
print(pew.model.4$formula)
```

```
## LIFE ~ PPEDUCAT + PPINCIMP + IDEO + PPGENDER
```

```r
print(m.4.dev)
```

```
## [1] 781.2531
```

```r
m.5.dev = valid.dev(pew.model.5, pew2.valid)
print(pew.model.5$formula)
```

```
## LIFE ~ PPEDUCAT + PPINCIMP + IDEO + PPGENDER + PPETHM
```

```r
print(m.5.dev)
```

```
## [1] 782.8132
```

```r
m.6.dev = valid.dev(pew.model.6, pew2.valid)
print(pew.model.6$formula)
```

```
## LIFE ~ PPEDUCAT + PPINCIMP + IDEO + PPGENDER + PPETHM + PPREG4
```

```r
print(m.6.dev)
```

```
## [1] 784.2374
```

```r
# Choose the model with the minimum deviance
devs = c(m.0.dev, m.1.dev, m.2.dev, m.3.dev, m.4.dev, m.5.dev, m.6.dev)
best.index = which.min(devs)
cat("The model with the minimum deviance was the model with", best.index-1, "predictors")
```

```
## The model with the minimum deviance was the model with 2 predictors
```

```r
best.model = pew.model.2
```

Based on the performance of these models on the validation set, which do you choose? (your answer here): LIFE ~ PPEDUCAT + PPINCIMP

Question 7 - 10 points

Please assess the performance of the model you chose in Question 6 as applied to the test data set. Please include a confusion matrix and compute accuracy, precision, recall, and F1 score for this model.

```r
# Your code here
# Predict on the best model
preds = predict(best.model, pew2.test, type="response")
# Turn the prediction probability into classifications
pred = ifelse(preds >= 0.5, 1, 0)

conf.mat = table(pred, pew2.test$LIFE)
conf.mat
```

```
##
## pred   0   1
##    0 179 134
##    1 126 138
```

```r
# Accuracy = (TP + TN) / (TP + TN + FP + FN)
acc = (conf.mat[1,1]+conf.mat[2,2]) / sum(conf.mat)
# Precision = TP / (TP + FP)
prec = (conf.mat[2,2]) / (conf.mat[2,2] + conf.mat[2,1])
# Recall = TP / (TP + FN)
recall = (conf.mat[2,2]) / (conf.mat[2,2] + conf.mat[1,2])
# F1 = (2*Prec*Recall) / (Prec + Recall)
f1 = (2*prec*recall) / (prec+recall)

cat("Accuracy:", acc, "\n")
```

```
## Accuracy: 0.5493934
```

```r
cat("Precision:", prec, "\n")
```

```
## Precision: 0.5227273
```

```r
cat("Recall:", recall, "\n")
```

```
## Recall: 0.5073529
```

```r
cat("F1 score:", f1, "\n")
```

```
## F1 score: 0.5149254
```