Assignment ON13

Elmer V Villanueva

Due at 3:55 PM on Monday 25 May 2020

Instructions

You must attempt all questions. All answers must be submitted on ICE. The *ONLY* submission format accepted is an RMD file.

Data

According to Fils-Aime et al. [1], epidemiologic surveys have found that alcoholism is the most common mental or substance abuse disorder among men in the United States. Fils-Aime and associates investigated the interrelationships of age at onset of excessive alcohol consumption, family history of alcoholism, psychiatric comorbidity, and cerebrospinal fluid (CSF) monoamine metabolite concentrations in abstinent, treatment-seeking alcoholics. Subjects were mostly white males classified as experiencing early (25 years or younger) or late (older than 25 years) onset of excessive alcohol consumption. Among the data collected were the following measurements on CSF tryptophan (TRYPT) and 5-hydroxyindoleacetic acid (5-HIAA) concentrations (pmol/ml). In this dataset, age of onset is zero if late or zero if early.

Regress the age of onset on the concentrations of both CSF metabolites, interpret the results and provide a diagnostic assessment of the model.

```
ID <- c(1:129)
HIAA <- c(57, 116, 81, 78, 206, 64, 123, 147, 102, 93,
          128, 69, 20, 66, 90, 103, 68, 81, 143, 121,
          149, 82, 100, 117, 41, 223, 96, 87, 96, 34,
          98, 86, 118, 84, 99, 114, 140, 74, 45, 51,
          99, 54, 93, 50, 118, 96, 49, 133, 105, 61,
          197, 87, 50, 109, 59, 107, 85, 156, 110, 81,
          53, 64, 57, 29, 34, 102, 51, 92, 104, 50,
          93, 146, 96, 112, 23, 109, 80, 111, 85, 131,
          58, 110, 80, 42, 80, 91, 102, 93, 98, 78,
          152, 108, 102, 122, 81, 81, 99, 73, 163, 109,
          90, 110, 48, 77, 67, 92, 86, 101, 88, 38,
          75, 35, 53, 77, 179, 151, 57, 45, 76, 46,
          98, 84, 119, 41, 40, 149, 116, 76, 96)
TRYPT <- c(3315, 2599, 3334, 2505, 3269, 3543, 3374, 2345, 2855, 2972,
           3904, 2564, 8832, 4894, 6017, 3143, 3729, 3150, 3955, 4288,
           3404, 2547, 3633, 3309, 3315, 3418, 2295, 3232, 3496, 2656,
           4318, 3510, 3613, 3117, 3496, 4612, 3051, 3067, 2782, 5034,
           2564, 4335, 2596, 2960, 3916, 2797, 3699, 2394, 2495, 2496,
           2123, 3320, 3117, 3308, 3280, 3151, 3955, 3126, 2913, 3786,
           3616, 3277, 2656, 4953, 4340, 3181, 2513, 2764, 3098, 2900,
           4125, 6081, 2972, 3962, 4894, 3543, 2622, 3012, 2685, 3059,
           3946, 3356, 3671, 4155, 1923, 3589, 3839, 2627, 3181, 4428,
           3303, 5386, 3282, 2754, 4321, 3386, 3344, 3789, 2131, 3030,
           4731, 4581, 3292, 4494, 3453, 3373, 3787, 3842, 2882, 2949,
           2248, 3203, 3248, 3455, 4521, 3240, 3905, 3642, 5233, 4150,
           2579, 3249, 3381, 4020, 4569, 3781, 2346, 3901, 3822)
ONSET <- c(1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1,
           0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
           1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1,
           1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0,
           1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0,
           1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1,
           1, 1, 0, 1, 1, 1, 1, 1, 1)
ONSET.F <- factor(ONSET, levels = c(0, 1), labels = c("Early", "Late"))
ALCOHOL <- data.frame(ID, HIAA, TRYPT, ONSET, ONSET.F)
str(ALCOHOL)
```

```
## 'data.frame':    129 obs. of  5 variables:
##  $ ID     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ HIAA   : num  57 116 81 78 206 64 123 147 102 93 ...
##  $ TRYPT  : num  3315 2599 3334 2505 3269 ...
##  $ ONSET  : num  1 0 1 0 0 1 0 1 1 1 ...
##  $ ONSET.F: Factor w/ 2 levels "Early","Late": 2 1 2 1 1 2 1 2 2 2 ...
```

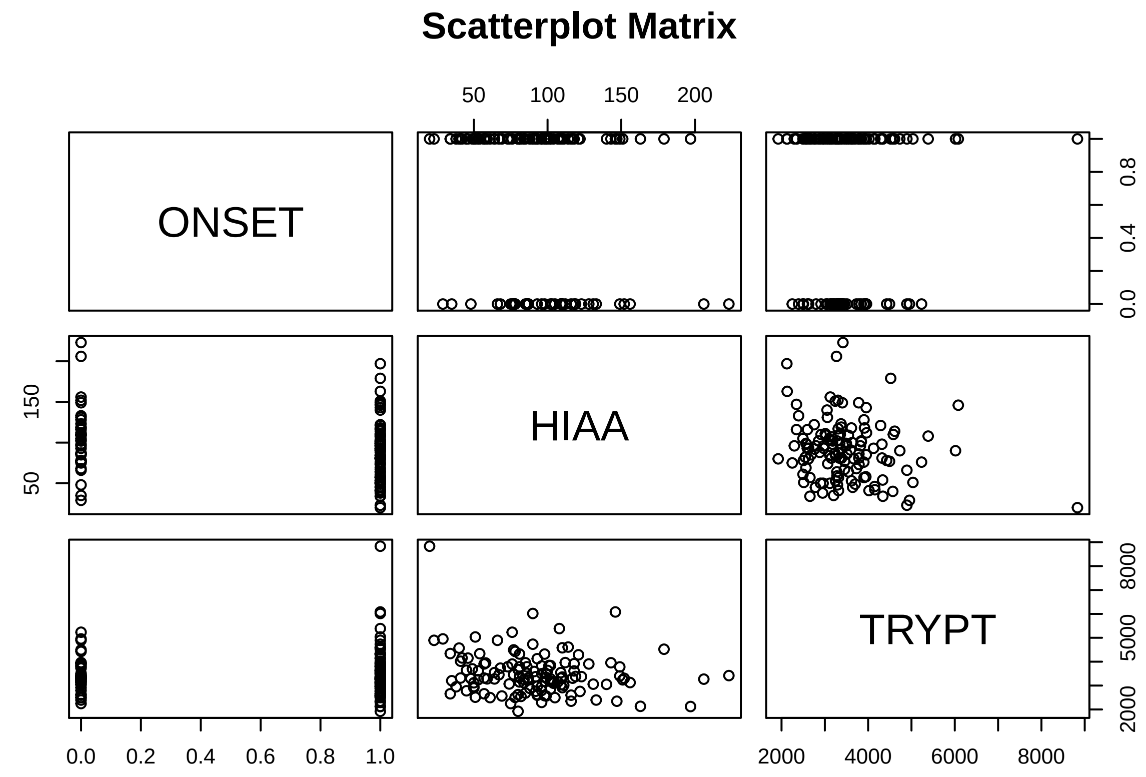
```
head(ALCOHOL)
```

```
##   ID HIAA TRYPT ONSET ONSET.F
## 1  1   57  3315     1    Late
## 2  2  116  2599     0   Early
## 3  3   81  3334     1    Late
## 4  4   78  2505     0   Early
## 5  5  206  3269     0   Early
## 6  6   64  3543     1    Late
```

Questions

1. (5 marks) Produce a properly-formatted scatterplot of the data.

```
pairs(~ONSET + HIAA + TRYPT, data = ALCOHOL,
      main = "Scatterplot Matrix")
```



2. (5 marks) Find the logistic regression equation describing the relationship among these variables. Produce the regression equation.

```
ALCOHOL.LM <- glm(ONSET.F ~ HIAA + TRYPT, data = ALCOHOL,
                  family = binomial(link = 'logit'))
summary(ALCOHOL.LM)
```

```
##
## Call:
## glm(formula = ONSET.F ~ HIAA + TRYPT, family = binomial(link = "logit"),
##     data = ALCOHOL)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9385  -1.3250   0.6931   0.8588   1.4165
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.076e+00  1.049e+00   1.979   0.0478 *
## HIAA        -1.336e-02  5.512e-03  -2.425   0.0153 *
## TRYPT        5.055e-06  2.330e-04   0.022   0.9827
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 158.11  on 128  degrees of freedom
## Residual deviance: 151.66  on 126  degrees of freedom
## AIC: 157.66
##
## Number of Fisher Scoring iterations: 4
```

The equation is $logit(ONSET) = 2.1 - (1.3 \times 10^{-2} \times HIAA) + (5.1 \times 10^{-4} \times TRYPT)$. Note that neither HIAA nor TRYPT are in unit values. HIAA is in 10's or 100's and TRYPT is in 1,000's. Thus, it might be more reasonable to report the following equation

$logit(ONSET) = 2.1 - 1.3 \times (100HIAA) + 0.51 \times (1000TRYPT)$

3. (10 marks) Interpret the point estimates of the estimated slope parameters.

```
exp(coef(ALCOHOL.LM))
```

```
## (Intercept)        HIAA       TRYPT
##   7.9695165   0.9867248   1.0000051
```

After controlling for TRYPT, every unit increase in HIAA results in a decrease in the odds of early onset alcoholism by about 1.3%. The odds to early onset alcoholism increases by 0.00051% for every unit increase in TRYPT after holding HIAA constant.

As noted above, unit increases in either HIAA or TRYPT are not very meaningful. Thus, you might want to report the results in this manner: After controlling for TRYPT, for every 100 units that HIAA increases, the odds of early onset alcoholism decrease by about $e^{-1.336} - 1 = 0.2629 - 1 = 73.71\%$. The odds to early onset alcoholism increase by $e^{0.5055} - 1 = 1.6578 - 1 = 65.78\%$ for every 1,000 units that TRYPT increases, after holding HIAA constant.

4. (10 marks) Report 95% confidence intervals of the estimated slope patameters.

```
confint(ALCOHOL.LM)
```

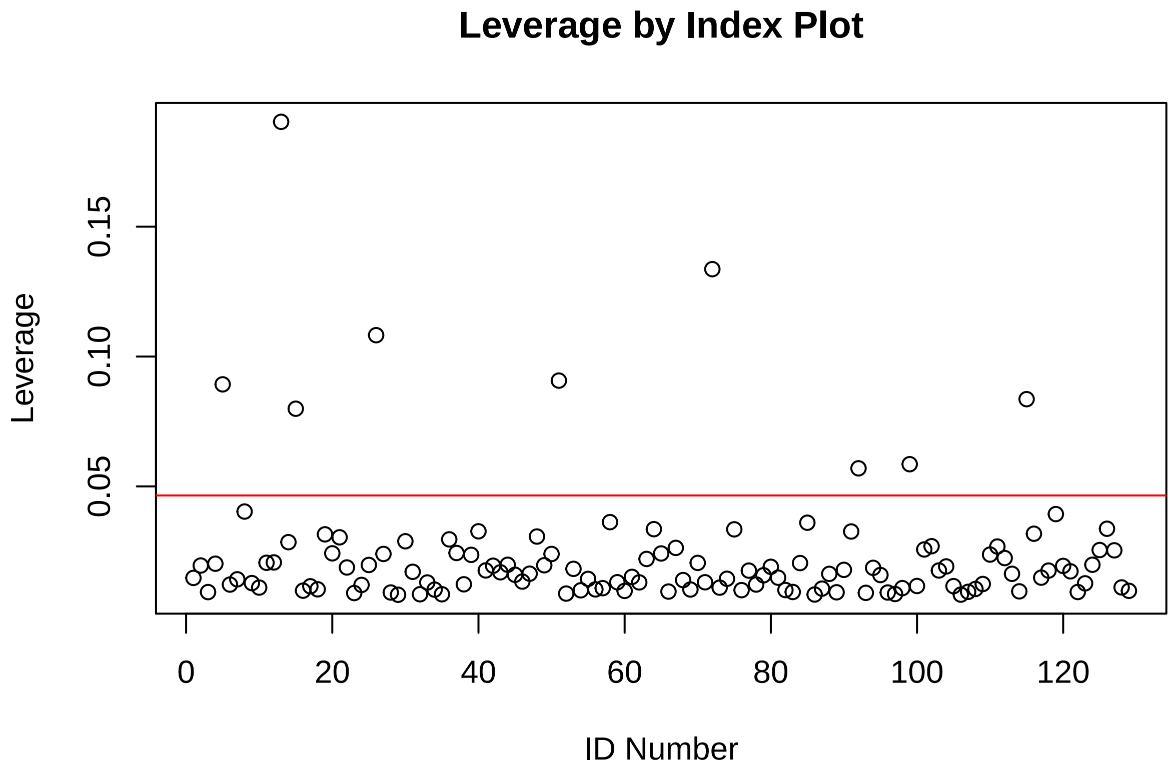
```
## Waiting for profiling to be done...

##                      2.5 %        97.5 %
## (Intercept)  0.0171940181  4.1671223415
## HIAA        -0.0246882900 -0.0028467457
## TRYPT       -0.0004330753  0.0004898807
```

The 95% CI for HIAA is (-0.0247, -0.0028). The 95% CI for TRYPT is $(-4.331 \times 10^{-4}, 4.899 \times 10^{-4})$.

5. (10 marks) Evaluate the presence of outliers in the independent variables using leverage values. What do you conclude?

```
ALCOHOL$HAT <- hatvalues(ALCOHOL.LM)
HAT.CUT <- 2 * (2 + 1)/ length(ID)
plot(HAT ~ ID, data = ALCOHOL,
     ylab = "Leverage",
     xlab = "ID Number",
     main = "Leverage by Index Plot")
abline(h = HAT.CUT, col = "red")
```



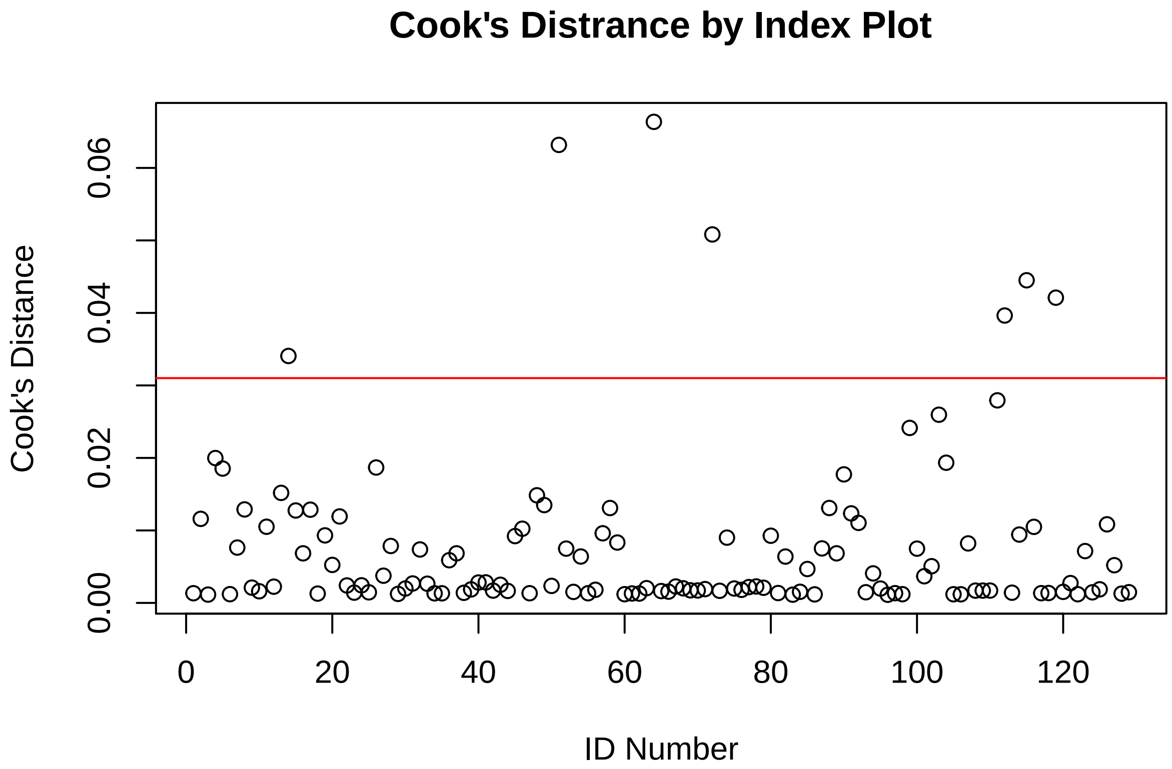
```
ALCOHOL[ALCOHOL$HAT > HAT.CUT,]
```

```
##       ID HIAA TRYPT ONSET ONSET.F        HAT
## 5      5  206  3269     0   Early 0.08928632
## 13    13   20  8832     1    Late 0.19036869
## 15    15   90  6017     1    Late 0.07990276
## 26    26  223  3418     0   Early 0.10822071
## 51    51  197  2123     1    Late 0.09072977
## 72    72  146  6081     1    Late 0.13365183
## 92    92  108  5386     1    Late 0.05695178
## 99    99  163  2131     1    Late 0.05854164
## 115 115  179  4521     1    Late 0.08358640
```

Nine out of 129 observations were noted to have high leverage values. Seven of the nine observations had hat values of about 0.10 or less. Observation 13 had a hat value of about twice this.

6. (10 marks) Evaluate the presence of influential observations using Cook's distance. What do you conclude?

```
ALCOHOL$COOK <- cooks.distance((ALCOHOL.LM))
COOK.CUT <- 4/length(ID)
plot(COOK ~ ID, data = ALCOHOL,
     ylab = "Cook's Distance",
     xlab = "ID Number",
     main = "Cook's Distrance by Index Plot")
abline(h = COOK.CUT, col = "red")
```



```
ALCOHOL[ALCOHOL$COOK > COOK.CUT,]
```

```
##      ID HIAA TRYPT ONSET ONSET.F        HAT       COOK
## 14   14   66  4894     0   Early 0.02851788 0.03405968
## 51   51  197  2123     1    Late 0.09072977 0.06317323
## 64   64   29  4953     0   Early 0.03352516 0.06635281
## 72   72  146  6081     1    Late 0.13365183 0.05082376
## 112 112   35  3203     0   Early 0.02240027 0.03964010
## 115 115  179  4521     1    Late 0.08358640 0.04450261
## 119 119   76  5233     0   Early 0.03933081 0.04210003
```

Seven of 129 observations had D values greater than the cut-off.

7. (10 marks) Report goodness of fit using the Hosmer-Lemeshow test and Nagelkerke's R^2. Interpret these values. What do you conclude?

```
if(!require(LogisticDx)){install.packages("LogisticDx")}
```

```
## Loading required package: LogisticDx

## Warning: package 'LogisticDx' was built under R version 3.6.3
```

```
library("LogisticDx")
gof(ALCOHOL.LM, g = 9, plotROC = FALSE)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

##        chiSq  df     pVal
## PrI  128.14 126 0.430097
## drI  151.66 126 0.059468 .
## PrG  128.14 126 0.430097
## drG  151.66 126 0.059468 .
## PrCT 128.14 126 0.430097
## drCT 151.66 126 0.059468 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                       val df     pVal
## HL chiSq          7.29693  7 0.398632
## mHL F             0.88449  8 0.531782
## OsRo Z            1.64750 NA 0.099455 .
## SstPgeq0.5 Z      0.67431 NA 0.500113
## SstPl0.5 Z        0.36138 NA 0.717817
## SstBoth chiSq     0.58529  2 0.746287
## SllPgeq0.5 chiSq 0.45507  1 0.499938
## SllPl0.5 chiSq   0.13918  1 0.709100
## SllBoth chiSq    1.87458  2 0.391689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

IMPORTANT: For this section, I needed to adjust the g option of `gof()`.

The Hosmer-Lemeshow test failed to reject the hypothesis that that model is a poor fit for the data.

```r
if(!require(rms)){install.packages("rms")}
```

```
## Loading required package: rms

## Warning: package 'rms' was built under R version 3.6.3

## Loading required package: Hmisc

## Warning: package 'Hmisc' was built under R version 3.6.3

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

## Loading required package: SparseM

##
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
##
##     backsolve
```

```r
library("rms")
ALCOHOL.LM.DX <- lrm(ONSET.F ~ HIAA + TRYPT)
print(ALCOHOL.LM.DX)
```

```
## Logistic Regression Model
##
##  lrm(formula = ONSET.F ~ HIAA + TRYPT)
##
##                       Model Likelihood     Discrimination    Rank Discrim.
##                          Ratio Test           Indexes           Indexes
##  Obs           129    LR chi2      6.45    R2       0.069    C       0.641
##   Early         39    d.f.            2    g        0.541    Dxy     0.283
##   Late          90    Pr(> chi2) 0.0397    gr       1.718    gamma   0.283
##  max |deriv| 5e-10                         gp       0.112    tau-a   0.120
##                                            Brier    0.201
##
##            Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept  2.0756 1.0487  1.98  0.0478
##  HIAA      -0.0134 0.0055 -2.42  0.0153
##  TRYPT      0.0000 0.0002  0.02  0.9827
##
```

The Nagelkirke pseudo-R squared value is 6.9%, meaning that less than 7% of variation in the dependent variable is explained by HIAA and TRYPT, and 93% of the variation is still unknown.

References

1. Fils-Aime ML, Eckardt MJ, George DT, Brown GL, Mefford I, Linnoila M. Early-onset alcoholics have lower cerebrospinal fluid 5-hydroxyindoleacetic acid levels than late-onset alcoholics. *Archives of General Psychiatry* 1996;53:211-216.

THE END