# Problem Set 4

PUBHLTH 490Z

Due Tuesday Dec 07 2021, 11:59PM EST

**Topics**

- Logistic Regression
- Data Analysis
- Intro to prediction/classification

**Logistic regression problem**

**Sleep Trouble.**

Using NHANES data, we will examine the increased odds of sleep trouble (`SleepTrouble`) in relation to age and gender. `SleepTrouble` is coded `1` if an individual has sleep trouble, and `0` if otherwise.

From this dataset, we will use the following variables to build a model to predict `SleepTrouble`.

- `Gender`
- `Age`
- `Education`
- `Poverty`
- `Work`
- `Depressed`

a) Using all the variables given above, build a logistic regression model with `SleepTrouble` as the outcome.

```
load("nhanes.samp.adult.Rdata")
#fit the full model with all predictors
```

b) Conduct a likelihood ratio test (LRT) to determine whether `Depression` is a statistically significant predictor of `SleepTrouble`. There are missing values (NA) in the `Depression` variable - make sure to handle these appropriately in the LRT test.

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.0.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.0.2
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
?lrtest
```

c) Using the variables given above to build a parsimonius, best fitting model using `Forward Selection`, use AIC as the criterion for selecting the best model.

```
#use the step() function
```

d) Interpret the `all` coefficients in best model selected in step c) above.

e) Refit the model in d) by setting the reference category for the `Depressed` variable as the level "Most". How do the coefficients of the `Depressed` variable change and what is the interpretation of these coefficients in terms of odds ratios?

```
?relevel
```

f) Use the R package `caret` to estimate the accuracy of the logistic regression model (classifier) in step (c), using 5-fold CV. Set the missing values of any predictors to the mean (if continuous) or to the mode (categorical).

**ICU: Probability of Survival.**

Patients admitted to an intensive care unit (ICU) are either extremely ill or considered to be at great risk of serious complications requiring the special technology and highly skilled care available in an ICU. There are no widely accepted criteria for distinguishing between patients who should be admitted to an ICU and those for whom admission to other hospital units would be more appropriate. Thus, among different ICU's, there are wide ranges in a patient's chance of survival. When studies are done to compare effectiveness of ICU care, therefore, it is critical to have a reliable means of assessing the comparability of the different patient populations.

One such strategy for doing so involves the use of statistical modeling to relate empirical data for many patient variables to outcomes of interest. The following dataset consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult ICU.[1] The major goal of the study was to develop a logistic regression model to predict the probability of survival to hospital discharge, and to study the risk factors associated with ICU mortality.[2]

The dataset `icu` contains the following variables:

- `vital.status`: whether the patient survived to hospital discharge (`lived`), or did not (`died`)
- `age`: age, measured in years
- `gender`: gender, either `male` or `female`
- `race`: either `white`, `black`, or `other`
- `service`: the type of service the patient needed upon ICU admission, either `medical` or `surgical`
- `cancer`: coded `yes` if cancer was part of the present problem, `no` if otherwise
- `renal`: coded `yes` if the patient had a history of chronic renal failure, `no` if otherwise
- `infect.prob`: `yes` if infection was probable, `no` if otherwise
- `cpr`: `yes` if CPR was administered prior to ICU admission, `no` if otherwise
- `sys`: measured in mm Hg. Typical systolic blood pressure ranges from 90 - 120 mm Hg.
- `hr`: measured in beats/min. Typical resting heart rate ranges from 60 - 100 bpm.
- `previous`: `yes` if previously admitted to an ICU within 6 months, `no` if otherwise
- `type`: type of admission, either `elective` or `emergency`
- `fracture`: coded `yes` if patient had a long bone, multiple, neck, single area, or hip fracture; `no` if otherwise
- `PO2`: oxygen partial pressure, measured in mm Hg. Normal arterial oxygen concentration is between 75-100 mm Hg, levels below 60 require supplemental oxygen.

[1] From Hosmer D.W., Lemeshow, S., and Sturdivant, R.X. *Applied Logistic Regression*. $3^{rd}$ ed., 2013. Accessible from `http://www.umass.edu/statdata/statdata/data/icu.txt`

[2] Lemeshow S., et al. Predicting the outcome of intensive care unit patients. *Journal of the American Statistical Association* 83.402 (1988): 348-356.

- `PH`: normal blood pH is typically between 7.35 and 7.45. Low blood pH is indicative of acidosis, which can have serious consequences.
- `PCO2`: carbon dioxide partial pressure, measured in mm Hg. Normal arterial $CO_2$ concentration is between 35-45 mm Hg. Values higher than 45 mm Hg is indicative of respiratory failure.
- `bicarb`: bicarbonate level, measured in mEq/L. Low bicarbonate levels are indicative of metabolic acidosis.
- `creat`: creatinine levels, measured in mg/dL. Typical ranges are 0.5 - 1.0 mg/dL. Elevated creatinine levels may be a sign of renal failure.
- `conscious`: level of consciousness at ICU admission, either `no coma/stupor`, `deep stupor`, or `coma`

a) Use numerical and graphical summaries to learn about the data.

   i. Describe the study population. Note that it is not necessary to incorporate each of the variables into the description; choose ones that you feel are relevant for giving an overall picture of the patients entering the ICU.

   ii. Explore the relationship between vital status and age, and between vital status and systolic blood pressure.

   iii. Explore the relationship between vital status and type of service needed, and between vital status and type of admission.

b) Conduct tests of association. For each test, summarize your results. (Note: For the purposes of this analysis, you may ignore any warning messages that `chisq.test()` returns.)

   i. Examine whether any demographic characteristics (i.e., age, gender, race) seem to be associated with vital status. (Hint: Visualizing the data such as by using `barplot(table())` may be useful for getting oriented.)

   ii. The information in columns 15-18 are collected as a part of an arterial blood gas (ABG) test. The test is used to check how well the lungs are able to move oxygen into the blood and remove carbon dioxide. Do any of the ABG measures seem to be associated with vital status?

   iii. Which factors seem to be associated with ICU mortality? Investigate the following: cancer-related admission, a history of chronic renal failure, probable infection, and presence of a bone fracture.

c) Create models.

   i. Based on the results of your tests of association conducted in part b), create a single model for predicting ICU mortality. Write the regression equation.

   ii. Calculate the relative odds of mortality for a patient for whom infection was probable versus a patient for whom infection was not probable, assuming that the other variables in the model are held constant.