**BIOS 6645**

**Predictive Analytics**

**Spring Semester 2018**

**Homework #1**

**Due: Friday February 16, 2018 by 5 PM to the Canvas Assignment Basket**

Submit your ***complete*** R code as an Appendix to your answers and include

only relevant R output with your answers to each part.

1. *Designs and scenarios*
   1. If a prediction model for risk of early dementia in high risk individuals is intended to estimate the burden of disease in subgroups or the actual risk for individuals, what metric would be the most important to maximize: discrimination or calibration? Why?

Estimation of the burden of a disease for a subgroup would be best optimized by focusing on calibration. If a model predicts 5% of the subgroup will get early dementia and actual early dementia rates are closer to 25% than the model isn’t very useful for determining disease burden.

* 1. If a prediction model for risk of early dementia in high risk individuals is intended to identify those at highest vs. lowest risk of disease, which metric would be the most important to maximize: discrimination or calibration? Why?

A model that identifies patients with highest vs lowest risk of disease would be best optimized by focusing on discrimination. A good discriminator will always flag individuals that actual have high risk with a higher risk score.

* 1. What would be the advantages and disadvantages of a nested case-control design vs. a full cohort for evaluating the predictive accuracy of a risk prediction model?

According to Wacholder (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2774111/>) the primary disadvantage of a nested case control design are the reduced power and precision due to sampling; additionally if sampling design has flaws than the analysis will be flawed. Full cohort design would not have the potential for reduced power or precision that nested case control design would have.

Nested case control requires less work in terms of less measurement of individuals/samples for evaluation and thus is an advantage over a full cohort study were every individual/sample must be evaluated.

* 1. What would be the advantages and disadvantages of a nested case-control design vs. a full cohort for evaluating calibration in an independent sample from a low risk population?

In terms of calibration, if the outcome/disease state is very rare (such as among a low risk population), the sampling method selected for nested case control design must ensure that those with the outcome/disease state are represented. If the outcome/disease is rare, a full cohort study design would likely be more appropriate as any sample selected from a low risk cohort would likely contain no individuals with the outcome in question.

1. *Simulation, linear regression, variance explained, predictions and predictive accuracy*

In this exercise you’ll create some simulated data from a linear model of a continuous outcome and will fit simple regression models to them. Make sure to set a seed prior to starting part (a) to ensure consistent results. Use base R and the rms package.

1. Using the rnorm()function, create a vector, x, containing 100 observations drawn from a N(0, 1) distribution. This represents a feature, X.
2. Using the rnorm()function, create a vector, eps, containing 100 observations drawn from a N(0, 0.25) distribution i.e. a normal distribution with mean zero and variance 0.25.
3. Using x and eps, generate a vector y according to the model

Y = -1 + 0.5X + ε

1. Create a scatterplot displaying the relationship between x and y. Comment on what you observe.
2. Fit a least squares linear model to predict y using x. Comment on the model obtained. How do  and  compare to  and ?
3. Now fit a polynomial regression model that predicts y using x and x2. Is there evidence that the quadratic term improves the model fit? Explain your answer.
4. Repeat (a)–(f) after modifying the data generation process in such a way that there is less noise in the data. Describe your results.
5. Produce two separate plots showing the prediction bands for individual predictions from each of the models. Comment on what you observe about the uncertainty of predictions across the two models. What would you expect if you had built in more (rather than less) noise in the data
6. Familiarize yourself with the rms package in R. Use the validate.ols function to produce an “optimism- corrected” R2 value for each of the two models above. Explain how this differs from producing a bootstrap distribution of the R2 values themselves.
7. *Logistic regression, calibration and discrimination*

An extensive data set was used to develop a prognostic model for 6-month unfavorable outcome (d.unfav) after moderate and severe TBI (d.gos = 3, 4). The strongest predictors were age, motor score and pupil reactivity. See “TBI Background Document.pdf” (in Canvas under Files -> Data) for the data dictionary. Note: Truncated systolic blood pressure (d.sysbt) is also included in the dataset but does not appear on the list in the document. The data are from Steyerberg and are in SPSS format. The dataset can be found in Canvas under Files -> Data -> TBI.sav.

* + 1. Using the rms package in R, fit a logistic regression model (binary outcome) with these 3 variables (age, d.motor and pupil.i) in the US data set (trial=75). First create 5 dummy variables for motor score and 2 for pupil reactivity. Be sure to investigate a potential nonlinear effect of age in the model.
    2. Obtain the predicted probability according to the model for each patient and plot these indicating values for favorable vs. unfavorable outcomes.
    3. What is the estimated fraction of unfavorable outcomes in the analyzed subset of patients? Verify that this is as expected for these data.
    4. Now, assess the discrimination of the model. Be sure to correct for the optimism of using the same data for fitting and assessing discrimination using the validate.lr function.
    5. Summarize the results in (a)-(d).
    6. What would be the next steps in terms of assessing generalizability of this model to patients in an international trial?
    7. EXTRA CREDIT: Carry out the next steps in (f) using the analogous international observations in the dataset (trial=74). Summarize the results for this test dataset.