Math 583 HW 9 2023 Due Wednesday, Nov. 8. Final: Wednesday, Dec. 13, 2:45-4:45.

```
A) 5.2): Logistic Regression Output for Problem 5.2
Response = nodal involvement, Terms = (acid size xray)
Label      Estimate  Std. Error     Est/SE    p-value
Constant  -3.57564   1.18002        -3.030     0.0024
acid       2.06294   1.26441         1.632     0.1028
size       1.75556   0.738348        2.378     0.0174
xray       2.06178   0.777103        2.653     0.0080
```

**5.2.** Following Collett (1999, p. 11), treatment for prostate cancer depends on whether the cancer has spread to the surrounding lymph nodes. Let the response variable = group $y$ = *nodal involvement* (0 for absence, 1 for presence). Let $x_1 = acid$ (serum acid phosphatase level), $x_2 = size$ (= tumor size: 0 for small, 1 for large) and $x_3 = xray$ (xray result: 0 for negative, 1 for positive). Assume the case to be classified has $\boldsymbol{x}$ with $x_1 = acid = 0.65$, $x_2 = 0$, and $x3 = 0$. Refer to the above output.

a) Find ESP for $\boldsymbol{x}$.

b) Is $\boldsymbol{x}$ classified in group 0 or group 1?

c) Find $\hat{\rho}(\boldsymbol{x})$.

Hint: $ESP = \hat{\alpha} + \sum_{i=1} x_i \hat{\beta}_i$ where $\hat{\alpha}$ corresponds to the constant and the $x_i$ are nontrivial predictors.

```
> out <- lda(x,group) #Problem 5.5
> 1-mean(predict(out,x)$class==group)
[1] 0.02
> out<-lda(x[,-c(1)],group)
> 1-mean(predict(out,x[,-c(1)])$class==group)
[1] 0.02
> out<-lda(x[,-c(1,2)],group)
> 1-mean(predict(out,x[,-c(1,2)])$class==group)
[1] 0.04
> out<-lda(x[,-c(1,3)],group)
> 1-mean(predict(out,x[,-c(1,3)])$class==group)
[1] 0.03333333
> out<-lda(x[,-c(1,4)],group)
> 1-mean(predict(out,x[,-c(1,4)])$class==group)
[1] 0.04666667
> out<-lda(x[,c(2,3,4)],group)
> 1-mean(predict(out,x[,c(2,3,4)])$class==group)
[1] 0.02
```

**B) 5.5.** The above output is for LDA on the famous iris data set. The variables are $x_1$ = sepal length, $x_2$ = sepal width, $x_3$ = petal length, and $x_4$ = petal width. These four predictors are in the $x$ data matrix. There are three groups corresponding to types of iris: setosa, versicolor, and virginica. Hint: See Example 5.2 and Problem C).

a) What is the AER using all 4 predictors?

b) Which variables, if any, can be deleted without increasing the AER in a)?

**C)** Read Example 5.2. The Wisseman et al. (1987) pottery data has 36 pottery shards of Roman earthware produced between second century B.C. and fourth century A.D. Often the pottery was stamped by the manufacturer. A chemical analysis was done for 20 chemicals (variables), and 28 cases were classified as Arrentine (group 1) or nonArrentine (group 2), while 8 cases were of questionable origin. So the training data has $n = 28$ and $p = 20$.

a) With $n = 28$ and $p = 20$, is this data set a high dimensional data set?

b) From Example 5.2, what was the AER using all 20 predictors?

c) From Example 5.2, what were the predictors that gave AER=0?

**D)** Suppose you are estimating the population median $\theta$ with the sample median $T = med(x)$.

actual data: 14, 3, 5, 12, 20, 10, 9: $med(x) = 10$,

a) Compute $T_1^*, T_2^*, T_3^*$, where $T_i^*$ is the sample median of the $i$th bootstrap sample.

bootstrap samples:

9, 10, 9, 12, 5, 14, 3

3, 9, 20, 10, 9, 5, 14

14, 12, 10, 20, 3, 3, 5

b) Now compute the bagging estimator $\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^*$ which is the sample mean of the $T_i^*$.

**E)** Consider high dimensional one sample tests for $H_0 : \boldsymbol{\mu} = \mathbf{0}$. The output gives the coverage and length of 95% "confidence intervals" for $\boldsymbol{\mu}^T\boldsymbol{\mu}$. The terms bcov and blen are for the $m$ out of $n$ bootstrap with $m \approx 2n/3$ where the shorth CI was applied to the bootstrap sample. The terms icov and ilen are for the $t$ CI based on the $W_i = \boldsymbol{x}_{2i-1}^T\boldsymbol{x}_{2i}$ for $i = 1, ..., k$ with $k \approx n/2$. The terms tcov and tlen are for the U-statistic CI using $\hat{\sigma}_W$ that is approximately the standard deviation of the $W_{ij} = \boldsymbol{x}_i^T\boldsymbol{x}_j$. The terms zcov and zlen are for U-statistic CI using $\hat{\sigma}_W = S_W =$ sample standard deviation of the $W_i$. For these two tests, the CI is only valid when $H_0$ is true so that $\boldsymbol{\mu}^T\boldsymbol{\mu} = 0$. The term cov gives the observed proportion of 100 runs where $H_0$ was rejected = power. When $H_0$ is true want cov $< 0.11$ for 100 runs. When $H_0$ is false want large power (the closer to one the better).

a) Copy and paste the $R$ commands for this problem, where $n = 100$, $p = 100$, and $\boldsymbol{\mu} = \mathbf{0}$ so $H_0$ is true. Which method has the longest length?

b) Copy and paste the $R$ commands for this problem, where $n = 100$, $p = 100$, and $\boldsymbol{\mu} = 0.1 \ \ \mathbf{1}$ so $H_0$ is false. How many methods had cov = power near 1?