ATTENTION!

Incorrect numerical answers unaccompanied by supporting reasoning will receive NO partial credit.

Correct numerical answers to difficult questions unaccompanied by supporting reasoning may not receive full credit.

SHOW YOUR WORK/EXPLAIN YOURSELF!
1. There are data on the UCI Machine Learning Repository due originally to I-Cheng Ye giving properties of a number of batches of concrete. Analyses for 103 batches are represented on R printouts at the end of this exam and pieces of output are also interspersed in the questions. Use both as appropriate.


5 pts a) Below is a graphic based on the "leaps" function regsubsets() for the \( n = 103 \) batches. **Which 3 predictors** seem to be most effective in predicting Strength? **What** fraction of the raw variability in Strength do they account for?

8 pts b) There are ANOVA results on page 8 of the output. Use them and **give the value of** and **degrees of freedom for an F statistic** for comparing the full model involving all predictors to a reduced model including only Cement, Fly.Ash, Water, and Coarse.Aggr as predictors.

\[ F = \text{___________} \quad \text{d.f.} = \text{_________}, \text{_________} \]
c) Below are some results (some cross-validation error sums of squares) from 10-fold cross-validation (run using the DAAG package).

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictors Included</th>
<th>CVSSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cement</td>
<td>5223</td>
</tr>
<tr>
<td>2</td>
<td>Cement, Fly.Ash</td>
<td>1511</td>
</tr>
<tr>
<td>3</td>
<td>Cement, Fly.Ash, Water</td>
<td>1148</td>
</tr>
</tbody>
</table>

**Do these results suggest** that the full model of b) is an improvement over the reduced model in b)? (Say "Yes" or "No.") **Which** model appears to be best on the basis of cross-validation? **Explain.**

**4 pts**

d) In addition to **Strength** the property "Slump" was measured for each of the 103 batches of concrete. (There is one `lm()` output for `Slump` on the printout on page 9. Right now we have no use for this output, but it is there to help fix concepts.) Below is some R code that was run. **What values** does "SlumpInd" take? **Which** of the 103 cases get **which** value of this new variable?

```r
> SlumpInd <- rep(0, 103)
> for (i in 1:103) if (Concrete$Slump[i] > 15) + SlumpInd[i] <- 1
```

**6 pts**
e) There is an initial logistic regression output on page 9 of the R output. **Does** it indicate that the data in hand are of "no help at all" explaining the probability that SlumpInd=1? **Explain** why or why not.
f) What about the series of three logistic regression outputs beginning on the top of page 10 indicates that of the corresponding models, the one with 2 predictor variables is detectably better than the one with a single predictor, and the one with 3 predictor variables is *not* detectably better than the one with 2? Give specific numerically-supported reasons.

g) On the basis of the 2-predictor logistic regression output on page 10, over the range of values tested by Prof. Ye, **which of the two variables** "Water" and "Slag" **seems most important** (is most clearly "statistically significant") in determining the probability that SlumpInd=1? **Explain.**

h) Here is some R code and the plot it produces. **Explain why** (in light of part d)) **you are not surprised** that the two lines on the plot are very similar. (\(a = y\)-intercept and \(b = \text{slope}\))

```r
> aa <- lm(Slump ~ Water + Slag, data=Concrete)
> aaa <- glm(SlumpInd ~ Water + Slag, data=Concrete, family="binomial")
> plot(Concrete$Water, Concrete$Slag, pch=SlumpInd+1)
> abline(a=((15-coef(aa)[1])/coef(aa)[3]), b=(-coef(aa)[2]/coef(aa)[3]), lwd=1)
> abline(a=((coef(aaa)[1])/coef(aaa)[3]), b=(-coef(aaa)[2]/coef(aaa)[3]), lwd=3)
```

![Plot of Concrete $Water$ vs. Concrete $Slag$ with two regression lines](image)
There were 83 batches with $\text{Fly.Ash} > 0$ represented in Yeh's data. For the rest of this problem, we will consider analysis of only those cases. Two models for these data are

$$\log(\text{Strength}) = \gamma_1 + \gamma_2 \log(\text{Cement}) + \gamma_3 \log(\text{Fly.Ash}) + \epsilon$$

and

$$\text{Strength} = \beta_1 (\text{Cement})^{\beta_2} (\text{Fly.Ash})^{\beta_3} + \epsilon$$

These two models have been fit to the data by least squares using `lm()` and `nls()` respectively, and the fits are represented on pages 11 and 12 of the output.

**6 pts**

i) The R code below produced the plot below it. (There are two different plotting symbols used on that plot: the open circle default and the cross that is "pc=3").

```r
> plot(Concrete1$Strength, exp(predict(a)), xlab="Strength", ylab="Predicted Strength")
> points(Concrete1$Strength, predict(aa), pc=3)
```

Least squares fitting of the two models are different procedures. **What aspects of this plot and the output however indicate** for this data set there is little practical difference?

**7 pts**

j) Give a $p$-value for testing the hypothesis $H_0 : \beta_2 = 0$ in the second model. **Interpret your result** in the context of the concrete testing.
k) There is code on page 12 that produced the contour below. (On this plot there are contours beginning at "elevation" 26 (on the lower left) and ending at "elevation" 58. (Cement, Fly.Ash) pairs in the data set are plotted with symbol size corresponding to the observed batch strength.) If one wishes produce concrete using Cement = 250 with Strength = 42, what value of Fly.Ash do you recommend? What about this scenario makes this determination "unsafe"?

2. Beginning on page 13 there is R code and output corresponding to a partially replicated $2^3$ factorial experiment due originally to R. Snee and appearing in Engineering Statistics by Hogg and Ledolter. This concerns the effects of three two-level factors indicated below

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-Polymer Type</td>
<td>Standard (−) vs New (But Expensive) (+)</td>
</tr>
<tr>
<td>B-Polymer Concentration</td>
<td>.01% (−) vs .04% (+)</td>
</tr>
<tr>
<td>C-Amount of an Additive</td>
<td>2 lb (−) vs 12 lb (+)</td>
</tr>
</tbody>
</table>

in a chemical process on the response variable

$y =$ percentage impurity produced

Use this R output to answer the rest of the questions on this Exam.

a) What is the value of $s_{pooled}$ for this data set? (Say where you found your value.)

b) Based on 2-sided 95% confidence limits what $+/-$ "margins of error" would you apply first to the 5 unreplicated single values in the data, and then to the sample means for the 3 samples of size 2?

single $y$'s margin __________________ $n = 2$ sample means $\bar{y}$ margin __________________
c) Consider first the data from only the low level of Factor C. Use your values from part b) and 4 sample means to below make an \((A \times B)\) interaction plot enhanced with error bars. Based on this plot alone do you think there are statistically detectable \(A \times B\) interactions (at the low level of C)? Explain.

\[\text{Interaction Plot}\]

d) Give 95% two-sided confidence limits for \(\alpha \beta \gamma_{222}\) (the three-factor interaction of A at its second level, B at is second level, and C at its second level). (Plug in completely, but you don't need to simplify.)

\[\text{Confidence Limits}\]

e) Considering which \(2^3\) factorial effects are statistically detectable, say in words what the data indicate about the effects of Factors A, B, and C on impurity.

f) Suppose that one assumes that only A and B main effects (and the overall mean) are important. Give the four corresponding predictions (\(\hat{y}\) values).

\[\hat{y}_{(1)} = \hat{y}_c : \quad \hat{y}_a = \hat{y}_{ac} : \]

\[\hat{y}_b = \hat{y}_{bc} : \quad \hat{y}_{ab} = \hat{y}_{abc} : \]
R Code and OutPut for Ye's Concrete Data

> summary(Concrete)

Cement           Slag           Fly.Ash          Water       Superplasticizer
Min.   :137.0   Min.   :  0.00   Min.   :  0.0   Min.   :160.0   Min.   : 4.40
1st Qu.:152.0   1st Qu.:  0.05   1st Qu.:115.5 1st Qu.:180.0 1st Qu.: 6.00
Median :248.0   Median :100.00   Median :164.0 Median :196.0 Median : 8.00
Mean   :229.9   Mean   : 77.97   Mean   :149.0   Mean   :197.2   Mean   : 8.54
3rd Qu.:303.9   3rd Qu.:125.00   3rd Qu.:235.9 3rd Qu.:209.5 3rd Qu.:10.00
Max.   :374.0   Max.   :193.00   Max.   :260.0 Max.   :240.0 Max.   :19.00

Coarse.Aggr       Fine.Aggr         Slump            Flow          Strength
Min.   : 708.0   Min.   :640.6   Min.   : 0.00   Min.   :20.00   Min.   :17.19
1st Qu.: 819.5   1st Qu.:684.5   1st Qu.:14.50 1st Qu.:38.50 1st Qu.:30.90
Median : 879.0   Median :742.7   Median :21.50 Median :54.00 Median :35.52
Mean   : 884.0   Mean   :739.6   Mean   :18.05  Mean   :49.61  Mean   :36.04
3rd Qu.: 952.8   3rd Qu.:788.0   3rd Qu.:24.00 3rd Qu.:63.75 3rd Qu.:41.20
Max.   :1049.9   Max.   :902.0   Max.   :29.00 Max.   :78.00 Max.   :58.53

> a<-regsubsets(Strength~Cement+Slag+Fly.Ash+Water+Superplasticizer+
+                  Coarse.Aggr+Fine.Aggr,nbest=1,data=Concrete)
> plot(a,scale="r2",main="Best R-Squares for Strength")

> anova(lm(Strength~Cement+Slag+Fly.Ash+Water+Superplasticizer+
          +Coarse.Aggr+Fine.Aggr,n,data=Concrete))

Analysis of Variance Table

Response: Strength

Df Sum Sq Mean Sq F value    Pr(>F)
Cement 1 1245.0  1245.0 182.9514 < 2.2e-16 ***
Slag   1  331.4   331.4  48.6954 3.999e-10 ***
Fly.Ash 1 3489.9  3489.9 512.8413 < 2.2e-16 ***
Water  1  344.1   344.1  50.5677 2.137e-10 ***
Superplasticizer 1  75.2   75.2   11.0473  0.001263 **
Coarse.Aggr 1  326.2   326.2  44.0074 1.809e-09 ***
Fine.Aggr  1  12.5    12.5   1.8386  0.178327
Residuals 95  646.5    6.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

> anova(lm(Strength~Cement+Fly.Ash+Water+Coarse.Aggr,data=Concrete))

Analysis of Variance Table

Response: Strength

Df Sum Sq Mean Sq F value    Pr(>F)
Cement       1 1245.0  1245.0 167.965 < 2.2e-16 ***
Fly.Ash      1 3589.9  3589.9 484.324 < 2.2e-16 ***
Water        1  379.1   379.1  51.150 1.557e-10 ***
Coarse.Aggr  1  326.2   326.2  44.007  0.001263 **
Residuals    98  726.4    7.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
> summary(lm(Slump~Water+Slag,data=Concrete))

Call:
  lm(formula = Slump ~ Water + Slag, data = Concrete)

Residuals:
     Min      1Q  Median      3Q     Max
-16.87   -5.33   2.49   5.43  11.36

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -18.0995    7.3139  -2.47   0.0150 *
Water         0.1989     0.0365   5.45  3.6e-07 ***
Slag         -0.0393     0.0122  -3.23   0.0017 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.44 on 100 degrees of freedom
Multiple R-squared:  0.291, Adjusted R-squared:  0.277
F-statistic: 20.6 on 2 and 100 DF,  p-value: 3.29e-08

> summary(glm(SlumpInd~Cement+Slag+Fly.Ash+Water+Superplasticizer +                +Coarse.Aggr+Fine.Aggr,data=Concrete,family="binomial"))

Call:
  glm(formula = SlumpInd ~ Cement + Slag + Fly.Ash + Water + Superplasticizer +
      Coarse.Aggr + Fine.Aggr, family = "binomial", data = Concrete)

Deviance Residuals:
     Min      1Q  Median      3Q     Max
-3.301  -0.670   0.470   0.708   1.449

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -66.72055   93.71547  -0.71     0.48
Cement               0.01403    0.02933   0.48     0.63
Slag                 0.00866    0.03946   0.22     0.83
Fly.Ash              0.01498    0.03067   0.49     0.63
Water                0.11612    0.09911   1.17     0.24
Superplasticizer    0.03437    0.16690   0.21     0.84
Coarse.Aggr          0.02241    0.03544   0.63     0.53
Fine.Aggr            0.02561    0.03779   0.68     0.50

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.449 on 102 degrees of freedom
    Residual deviance: 93.579 on 95 degrees of freedom
    AIC: 109.6

Number of Fisher Scoring iterations: 5
> summary(glm(SlumpInd~Water, data=Concrete, family="binomial"))

Call:
glm(formula = SlumpInd ~ Water, family = "binomial", data = Concrete)

Deviance Residuals:
  Min      1Q  Median      3Q     Max
-2.581  -1.024   0.569   0.796   1.314

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.73728    2.69023   -3.25  0.0012 **
Water        0.05011     0.01419    3.54  0.0004 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 122.45  on 102  degrees of freedom
Residual deviance: 106.54  on 101  degrees of freedom
AIC: 110.5

Number of Fisher Scoring iterations: 4

> summary(glm(SlumpInd~Water+Slag, data=Concrete, family="binomial"))

Call:
glm(formula = SlumpInd ~ Water + Slag, family = "binomial", data = Concrete)

Deviance Residuals:
  Min      1Q  Median      3Q     Max
-3.142  -0.759   0.501   0.727   1.404

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.15318    2.93573   -3.12  0.00182 **
Water        0.05867    0.01589    3.69  0.00022 ***
Slag        -0.01408    0.00472   -2.98  0.00284 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 122.45  on 102  degrees of freedom
Residual deviance:  96.05  on 100  degrees of freedom
AIC: 102.1

Number of Fisher Scoring iterations: 5

> summary(glm(SlumpInd~Water+Slag+Fly.Ash, data=Concrete, family="binomial"))

Call:
glm(formula = SlumpInd ~ Water + Slag + Fly.Ash, family = "binomial", data = Concrete)

Deviance Residuals:
  Min      1Q  Median      3Q     Max
-3.042  -0.736   0.494   0.697   1.388
Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | -7.94501 | 3.28786 | -2.42 0.01567 * |
| Water | 0.05471 | 0.01649 | 3.32 0.00091 *** |
| Slag | -0.01499 | 0.00488 | -3.07 0.00211 ** |
| Fly.Ash | -0.00250 | 0.00330 | -0.76 0.44736 |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 122.449 on 102 degrees of freedom
Residual deviance: 95.467 on 99 degrees of freedom
AIC: 103.5

Number of Fisher Scoring iterations: 5

> Concrete1<-data.frame(Concrete[Concrete$Fly.Ash > 0,])
> summary(Concrete1)

> a<-lm(log(Strength)~log(Cement)+log(Fly.Ash),data=Concrete1)
> summary(a)
```r
c1<-exp(coef(a)[1])
c2<-coef(a)[2]
c3<-coef(a)[3]

aa<-nls(Strength ~ b1*(Cement^b2)*(Fly.Ash^b3),data=Concrete1,
+     start=list(b1=c1,b2=c2,b3=c3),trace=T)

> summary(aa)

Formula: Strength ~ b1 * (Cement^b2) * (Fly.Ash^b3)

Parameters:

    Estimate Std. Error t value Pr(>|t|)
  b1    0.5248     0.1906    2.75   0.0073 **
  b2    0.4992     0.0363   13.74  < 2e-16 ***
  b3    0.3111     0.0454    6.85  1.4e-09 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.96 on 80 degrees of freedom

Number of iterations to convergence: 4
Achieved convergence tolerance: 1.49e-07

> confint(aa)

Waiting for profiling to be done...

1065 :  0.499 0.311
1324 :  0.518 0.336
.
.
.
1402 :  0.201 0.543
1402 :  0.201 0.543
  2.5% 97.5%
  b1 0.256 1.074
  b2 0.427 0.571
  b3 0.222 0.401

> Fit<- function(cement,flyash) {
+   coef(aa)[1]*(cement^coef(aa)[2])*(flyash^coef(aa)[3])
+ }
>
> cement<-seq(130,370,2)
fly<-seq(95,265,2)
FittedStrength<-outer(cement,fly,FUN=Fit)
> contour(cement,fly,FittedStrength,levels=seq(26,58,2),xlab="Cement",
+   ylab="Fly Ash",main="Predicted Strength",cex=2,lw=2)
> grid(col="black")
> points(Concrete1$Cement,Concrete1$Fly.Ash,pch=1,
+   cex=Concrete1$Strength/15)```
R Code and OutPut for Snee's Impurity Data

```r
> polytype <- c(1, 2, 1, 2, 1, 1, 2, 1, 2, 1)
> polyconc <- c(1, 1, 2, 1, 2, 1, 1, 2, 1, 2)
> addamt <- c(1, 1, 1, 2, 2, 2, 2, 2, 2, 2)
> typeindex <- c(1, 2, 3, 4, 5, 6, 7, 8)
> impurity <- c(1.0, 1.0, 1.2, .2, .5, .9, .7, 1.1, .2, .3, .5)
> polytype <- as.factor(polytype)
> polyconc <- as.factor(polyconc)
> addamt <- as.factor(addamt)
> typeindex <- as.factor(typeindex)
>
> Snee2 <- data.frame(typeindex, polytype, polyconc, addamt, impurity)
>
> aggregate(Snee2$impurity, by = list(Snee2$typeindex), mean)

<table>
<thead>
<tr>
<th>Group.1</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>2.10</td>
</tr>
<tr>
<td>3</td>
<td>3.20</td>
</tr>
<tr>
<td>4</td>
<td>4.50</td>
</tr>
<tr>
<td>5</td>
<td>5.80</td>
</tr>
<tr>
<td>6</td>
<td>6.10</td>
</tr>
<tr>
<td>7</td>
<td>7.25</td>
</tr>
<tr>
<td>8</td>
<td>8.50</td>
</tr>
</tbody>
</table>

> aggregate(Snee2$impurity, by = list(Snee2$typeindex), sd)

<table>
<thead>
<tr>
<th>Group.1</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>0.14142136</td>
</tr>
<tr>
<td>3</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>NA</td>
</tr>
<tr>
<td>5</td>
<td>0.14142136</td>
</tr>
<tr>
<td>6</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>0.07071068</td>
</tr>
<tr>
<td>8</td>
<td>NA</td>
</tr>
</tbody>
</table>
>
> options(contrasts = rep("contr.sum", 2))
>
> summary(lm(impurity ~ polytype * polyconc * addamt, data = Snee2))
```

Call:
```
lm(formula = impurity ~ polytype * polyconc * addamt, data = Snee2)
```

Residuals:
```
  Min  1Q Median  3Q Max
-1.00 0.00  0.00  0.00  2.00
```

Coefficients:
```
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.68125    0.03903  17.454  0.00041 ***
polytype1                -0.11875    0.03903  -3.042  0.05576 .
polyconc1                 0.31875    0.03903   8.167  0.00384 **
addamt1                   0.01875    0.03903   0.480  0.66381
polytype1:polyconc1      0.01875    0.03903   0.480  0.66381
polytype1:addamt1        0.01875    0.03903   0.480  0.66381
polyconc1:addamt1        0.03125    0.03903   0.801  0.48188
polytype1:polyconc1:addamt1 0.03125    0.03903   0.801  0.48188
...                           NA         NA         NA         NA
```

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.1225 on 3 degrees of freedom
Multiple R-squared: 0.9671, Adjusted R-squared: 0.8904
F-statistic: 12.61 on 7 and 3 DF, p-value: 0.0308

> A<-2*(as.numeric(polytype)-1.5)
> B<-2*(as.numeric(polyconc)-1.5)
> C<-2*(as.numeric(addamt)-1.5)
>
> summary(lm(impurity~A+B+C+(A*B)+(A*C)+(B*C)+(A*B*C)))

Call:
    lm(formula = impurity ~ A + B + C + (A * B) + (A * C) + (B * C) + (A * B * C))

Residuals:
     1       2       3       4       5       6       7       8       9      10      11
2.079e-17 -1.000e-01  1.000e-01 -4.723e-17  8.286e-18  1.000e-01 -1.000e-01  1.522e-17 -5.000e-02  5.000e-02  2.216e-17

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.68125    0.03903  17.454  0.00041 ***
A                        0.11875    0.03903   3.042  0.05576 .
B                        0.31875    0.03903  8.167  0.00384 **
C                        0.01875    0.03903  0.480  0.66381
A:B                      0.01875    0.03903  0.480  0.66381
A:C                      0.01875    0.03903  0.480  0.66381
B:C                      0.03125    0.03903  0.801  0.48188
A:B:C                    0.03125    0.03903  0.801  0.48188
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1

Residual standard error: 0.1225 on 3 degrees of freedom
Multiple R-squared: 0.9671, Adjusted R-squared: 0.8904
F-statistic: 12.61 on 7 and 3 DF, p-value: 0.0308

> summary(lm(impurity~A+B))

Call:
    lm(formula = impurity ~ A + B)

Residuals:
    Min     1Q Median     3Q    Max
-0.15926 -0.04074  0.01111  0.05000  0.14074

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.67407    0.02928  23.021 1.35e-08 ***
A                        0.12407    0.02928   4.237  0.00285 **
B                        0.30926    0.02928  10.562 5.64e-06 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1

Residual standard error: 0.09623 on 8 degrees of freedom
Multiple R-squared: 0.9459, Adjusted R-squared: 0.9324
F-statistic: 69.93 on 2 and 8 DF, p-value: 8.569e-06