Midterm EPI 853B

Name: _____

The exam is strictly individual. You can consult any materials you like but you cannot discuss your exam with anyone other than the instructor.

You have 80 to complete the exam.

For each question report in this document your results (in a clean and readable manner) and conclusions.

Provide a separate ASCII file with your code with appropriate comments indicating which question the code pertains to.

At the end of the exam e-mail me (gustavoc@msu.edu) both documents.

For all the questions of the exam you will use the Gout data set. A binary version of the data set is available at

https://github.com/gdlc/EPI853B/blob/master/gout.RData

Once you download this data, you can get into the R-environment using

```
load('~/GitHub/EPI853B/gout.RData')
```

Problem 1. Ordinary Least Squares Regression using lm

1.1 Regress Systolic blood pressure (SBP) on sex, race and age, report your results (model R-sq, estimates, SE, p-values) and summarize your findings.

Results

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 82.88874    4.74824  17.457  < 2e-16 ***
SexM         3.74366    0.99649   3.757  0.00018 ***
RaceW       -9.32869    1.17707  -7.925 5.13e-15 ***
Age          0.81506    0.08626   9.449  < 2e-16 ***
---
Multiple R-squared:  0.1128,   Adjusted R-squared:  0.1106
```

Findings

- All predictors have estimated effects that were significantly different from zero. Male have in average 3.74 more units of SBP than female, and whites have lower SBP than African Americans (9.32 units less for whites). We estimate that SBP increases by 0.82 units per year of age within the ranges of ages tested.
- However, Sex, Race and age only explain 11.3% of the variance of SBP (see R-sq).

1.2 Report the estimated expected SBP for each of the rows of the table.

Predictors			Predictions
Age	Sex	Race	
50	M	W	118.06
50	M	B	127.39
50	F	W	114.31
50	F	B	123.64
65	M	W	130.28
65	M	B	139.61
65	F	W	126.54
65	F	B	135.87

2. Compute and report OLS estimates and SE for the regression of problem 1 using matrix operations.

Report your results here

```
            Estimate   SE
(Intercept) 82.8887 4.7482
SexM         3.7437 0.9965
RaceW       -9.3287 1.1771
Age          0.8151 0.0863
```

3. Logistic regression of Gout (Yes/No) on Sex, Race and Serum Urate

3.1. Fit the model using glm(), report your results and summarize your findings

Results

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.8878     0.3219 -12.076  < 2e-16 ***
SexM          0.2579     0.1434   1.799  0.07210 .
RaceW        -0.4823     0.1581  -3.050  0.00229 **
UricAcid      0.5936     0.0466  12.738  < 2e-16 ***
```

Findings

Serum urate has a highly significant effect on the risk of developing gout, higher serum urate increases the risk of developing gout. B have higher risk of developing Gout than W (the odds for W/B is .61, that is about 40% less risk for W than for B). Finally, although male had in the sample higher incidence of Gout than females, the difference in risk between male and female is not statistically different than zero at the .05 significance level.

3.2. Fit the same regression via maximum likelihood using the optim() procedure. Reports the results obtained and comment on similarities/differences with those reported in 3.1

Results

```
              GLM     OPTIM
(Intercept) -3.8878 -3.8905
SexM         0.2579  0.2573
RaceW       -0.4823 -0.4813
UricAcid     0.5936  0.5940
```

Comments

We got similar estimates, with differences in the 2nd (intercept) and 3rd or higher order (regression coefficients) decimal place.

3.3. Did your ML procedure of 3.2 converged? How do you know that?

Converged? Yes

How do you know?

```
fm2$convergence
[1] 0

Which according to help(optim) corresponds to successful
convergence.
```

3.4. Estimate the SE of estimates of coefficients using 3000 Bootstrap samples.
Report your results and compare them with those reported by glm (3.1).

Results

```
(Intercept) 0.3219 0.3429
SexM        0.1434 0.1498
RaceW       0.1581 0.1616
UricAcid    0.0466 0.0489
```

Comments

Very similar, but not exactly equal SEs. The GLM SEs are based on the asymptotic variance of the ML estimates (based on the inverse of Fisher Observed Information). Bootstrap estimates do not reside on the same asymptotic argument.

Bonus Questions

1.2.bonus

Expand the table by adding estimates of SE of the estimated conditional expectation (explain your method and report results)

Your method:
Let W denote the incidence matrix for predictions, then the variance-covariance matrix of the estimated conditional expectation is given by

$$Var\left(W\hat{\beta}\right) = WVar\left(\hat{\beta}\right)W' = W(X'X)^{-1}W'\sigma_{\varepsilon}^{2}$$

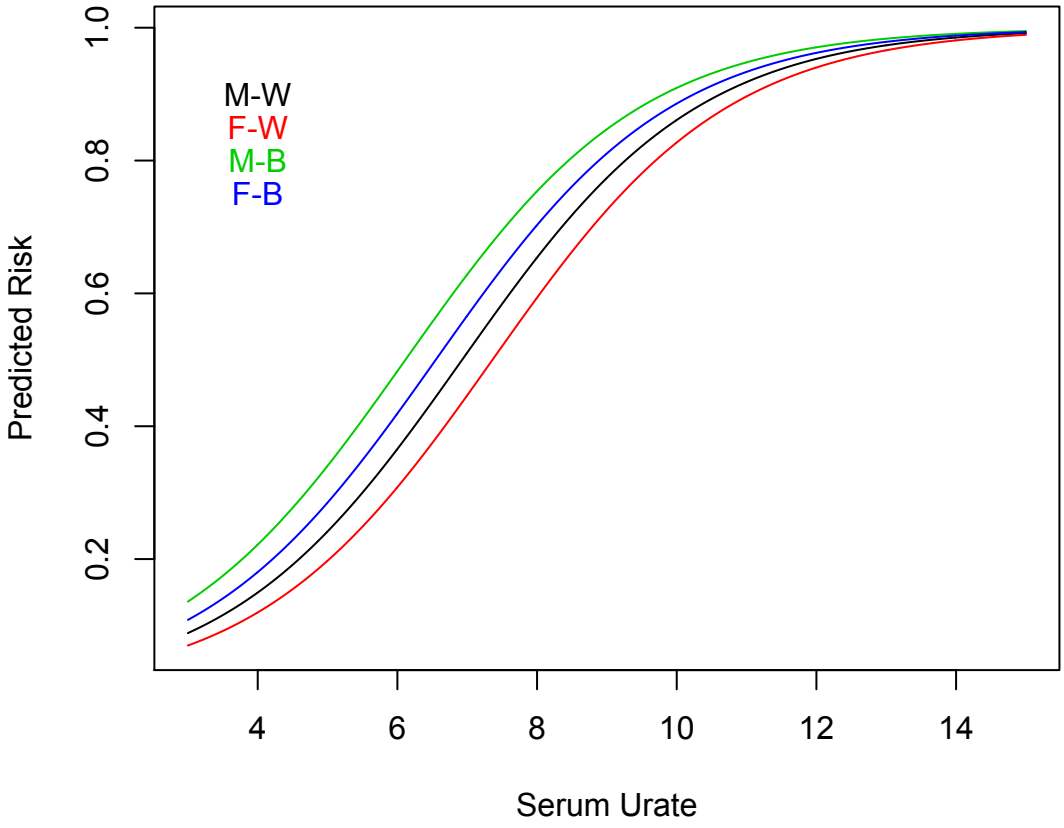
The square-root of the diagonal entries of the matrix described above gives the estimated SEs.

Predictors			Predictions SE and 95% CI for the estimated function yHat SE(yHat) Lower Upper
Age	Sex	Race	
50	M	W	118.0567 0.05154198 117.9557 118.1578
50	M	B	127.3854 0.07473281 127.2389 127.5319
50	F	W	114.3131 0.05450795 114.2062 114.4199
50	F	B	123.6418 0.07158653 123.5014 123.7821
65	M	W	130.2826 0.06979163 130.1458 130.4194
65	M	B	139.6113 0.09301547 139.4290 139.7936
65	F	W	126.5390 0.07439929 126.3931 126.6848
65	F	B	135.8677 0.09241946 135.6865 136.0488

3.bouns

With the model fitted in 1.1 plot curves for risk of developing Gout versus serum urate (for serum urate levels from 3 to 15) by sex-race combinations (i.e., 4 curves, one for each race-by-sex combinations).

Results


M-W
F-W
M-B
F-B
Predicted Risk
0.2
0.4
0.6
0.8
1.0
4
6
8
10
12
14
Serum Urate