DATA 11900
Seniors' final exam, Spring 2022
May 24, 2022 9:30am - 10:50am

Name: ____Student ID: ________________

1. Please **PRINT** your name and student ID in the above space.

2. This is a closed book, closed-notes examination.
You can have a single two-sided page of notes that you may refer to.
You may use calculators or calculator apps.

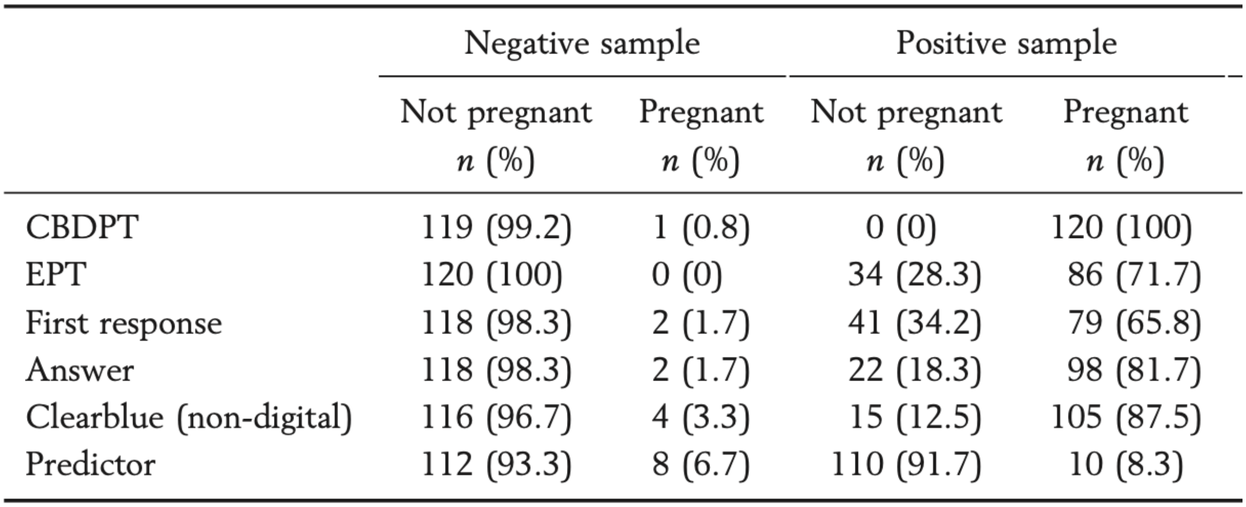
3. Please provide the answers in the space provided. The spaces provided are sufficient for short and clear explanations or justifications.

Problem	Total Points	Received Points
Home pregnancy test accuracy	13	
Covid statistics by county	20	
function	7	
Portuguese math grades	23	
regularization	7	

Tomlinson , Marshall & Ellis   Comparison of accuracy and certainty of results of six home pregnancy tests available over-the-counter Current Medical Research and Opinion v. 24, NO. 6, 2008, 1645–1649   doi:10.1185/03007990802120572

report distributing the urine of pregnant and non-pregnant women and testing the readings of 1400 over-the-counter urine pregnancy tests.

Table 2. Volunteer-recorded results of pregnancy tests performed on negative (hCG 0 IU/l) and positive (hCG 25 IU/l) urine samples using each of the home pregnancy tests (n = 120)



	Negative sample		Positive sample	
	Not pregnant *n* (%)	Pregnant *n* (%)	Not pregnant *n* (%)	Pregnant *n* (%)
CBDPT	119 (99.2)	1 (0.8)	0 (0)	120 (100)
EPT	120 (100)	0 (0)	34 (28.3)	86 (71.7)
First response	118 (98.3)	2 (1.7)	41 (34.2)	79 (65.8)
Answer	118 (98.3)	2 (1.7)	22 (18.3)	98 (81.7)
Clearblue (non-digital)	116 (96.7)	4 (3.3)	15 (12.5)	105 (87.5)
Predictor	112 (93.3)	8 (6.7)	110 (91.7)	10 (8.3)

Declaration of interest: This study was funded by SPD Swiss Precision Diagnostics, manufacturer of the Clearblue Digital Pregnancy Test and the Clearblue non-digital pregnancy test. CT, JM and JEE are all employees of SPD Swiss Precision Diagnostics (formerly known as Inverness Medical Innovation – Unipath Division).

1a. (4pts) Fill out the confusion matrix for the Clearblue home pregnancy test. Note the study design called for balanced true negative and true positive samples.

	True negative	True positive
Test negative	TN	FN
Test positive	FP	TP

	True negative	True positive
Test negative		
Test positive		

1b. (4pts) Find the crude accuracy of the Answer test assuming that 10% of test takers are pregnant.

1c. (5pts) Either positive predictive value or negative predictive value is a statistic of keen interest to the users of tests, depending on the result of the test. Calculate the negative predictive value and positive predictive value of the Clearblue home pregnancy test in this study.

COVID statistics by county

There was once great interest in collecting and reporting COVID cases and deaths by geography and time. The following tables are from The New York Times' Coronavirus (Covid-19) Data in the United States (https://github.com/nytimes/covid-19-data) and CDC Provisional COVID-19 Deaths by County, and Race and Hispanic Origin (https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-County-and-Race-and/k8wy-p9cg). The CDC data gives cumulative case numbers and cumulative deaths by date and county; the NYT database includes population estimates from the 2020 census.

Here are two sets of five lines from the countydeaths data table:

index	date	county	state	fips	cases	deaths
0	2020-01-21	Snohomish	Washington	53061	1	0.0
1	2020-01-22	Snohomish	Washington	53061	1	0.0
2	2020-01-23	Snohomish	Washington	53061	1	0.0
3	2020-01-24	Cook	Illinois	17031	1	0.0
4	2020-01-24	Snohomish	Washington	53061	1	0.0

index	date	county	state	fips	cases	deaths
1334021	2021-05-19	Livingston	Michigan	26093	16542	182.0
662971	2020-10-24	Portage	Ohio	39133	1687	68.0
1261995	2021-04-27	Stark	Illinois	17175	619	23.0
742329	2020-11-18	Washington	Florida	12133	1417	24.0
747966	2020-11-19	Maverick	Texas	48323	4690	169.0

2a. (4pts) Comment on the difference between the two samples above.

Suppose you have loaded the following two tables into SQL tables countydeaths and population. Here are five random rows from each:

index	date	county	state	fips	cases	deaths
1334021	2021-05-19	Livingston	Michigan	26093	16542	182.0
662971	2020-10-24	Portage	Ohio	39133	1687	68.0
1261995	2021-04-27	Stark	Illinois	17175	619	23.0
742329	2020-11-18	Washington	Florida	12133	1417	24.0
747966	2020-11-19	Maverick	Texas	48323	4690	169.0

countydeaths

index	us_state_fips	us_county_fips	population	region	subregion
1940	37	37099	42256	North Carolina	Jackson
699	17	17201	286174	Illinois	Winnebago
1029	21	21065	14313	Kentucky	Estill
902	20	20021	20331	Kansas	Cherokee
483	13	13187	31951	Georgia	Lumpkin

population

2b. (3pts) Write an SQL expression that evaluates the total population of all counties (better be about 331 Million).

SELECT

FROM

WHERE

2c. (3pts) Write an SQL expression that counts the number of counties that had more than 100 cases by 2021-06-01.

SELECT

FROM

WHERE

2d. (3pts) Write an SQL expression that counts the total population of all the counties with more than 100 cases by 2021-06-01.

SELECT

FROM

WHERE

2e. (3pts) Write an SQL expression that gives the names and counts of counties that occur in more than 9 states. (Hint: this is easier done with the population table than the countydeaths table, since the population table contains each county only once).

SELECT

FROM

WHERE

GROUP BY

HAVING

2f. (4pts) If the deaths column in countydeaths had values that were not-a-number, what would you do?

____

____

____

____

The sigmoid function is the function that converts from log-odds (using the natural logarithm) to probability.

3a. (3pts) Write the expression for the sigmoid function of z.

3b. (4pts) When is the sigmoid function useful in data analysis? (What is it used for?)

Portuguese High School Math grades

Cortez and Silva published a database of the math grades of a cohort of 395 students at two Portuguese high schools along with the results of brief surveys. Excerpts from the data dictionary are here:

# Attributes for student-math dataset:

sex - student's sex (binary: "F" - female or "M" - male)
age - student's age (numeric: from 15 to 22)
Medu - mother's education (numeric: 0 - none, to 4 – higher education)
Fedu - father's education (numeric: 0 - none, to 4 – higher education)
famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
absences - number of school absences (numeric: from 0 to 93)

# these grades are related with the course subject, Math or Portuguese:
G1 - first period grade (numeric: from 0 to 20)
G2 - second period grade (numeric: from 0 to 20)
G3 - final grade (numeric: from 0 to 20, output target)
And some summary statistics:

```
mathnum.mean()

Medu         2.749
Fedu         2.522
age         16.696
absences     5.709
Walc         2.291
famrel       3.944
G1          10.909
G2          10.714
G3          10.415
```

```
mathnum.std()

Medu        1.095
Fedu        1.088
age         1.276
absences    8.003
Walc        1.288
famrel      0.897
G1          3.319
G2          3.762
G3          4.581
```

Here is a table of Pearson's correlation coefficients between each pair of columns:

```
mathnum.corr()
```

	Medu	**Fedu**	**sex**	**age**	**absences**	**Walc**	**famrel**	**G1**	**G2**	**G3**
Medu	1.000	0.623	0.078	-0.164	0.100	-0.047	-0.004	0.205	0.216	0.217
Fedu	0.623	1.000	0.035	-0.163	0.024	-0.013	-0.001	0.190	0.165	0.152
sex	0.078	0.035	1.000	-0.029	-0.067	0.274	0.059	0.092	0.091	0.103
age	-0.164	-0.163	-0.029	1.000	0.175	0.117	0.054	-0.064	-0.143	-0.162
absences	0.100	0.024	-0.067	0.175	1.000	0.136	-0.044	-0.031	-0.032	0.034
Walc	-0.047	-0.013	0.274	0.117	0.136	1.000	-0.113	-0.126	-0.085	-0.052
famrel	-0.004	-0.001	0.059	0.054	-0.044	-0.113	1.000	0.022	-0.018	0.051
G1	0.205	0.190	0.092	-0.064	-0.031	-0.126	0.022	1.000	0.852	0.801
G2	0.216	0.165	0.091	-0.143	-0.032	-0.085	-0.018	0.852	1.000	0.905
G3	0.217	0.152	0.103	-0.162	0.034	-0.052	0.051	0.801	0.905	1.000

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, 2008, ISBN 978-9077381-39-7. https://archive.ics.uci.edu/ml/datasets/student+performance

4a: (10 points) Find the simple-least squares equation for G3 as a function of G2. This entails finding the coefficient and the intercept.

(this page intentionally left blank)

Multiple linear regression is performed to estimate G3 based on seven student-describing columns and gives the following coefficients:

	coefficient
Medu	0.691
Fedu	0.092
sex	0.944
age	-0.481
absences	0.033
Walc	-0.209
famrel	0.251
intercept_	15.162

4b: (3pts) Write the linear regression model equation for G3 for the above model.

____

____

4c: (4pts) The correlation between mother's education and father's education (Medu and Fedu) and final grade G3 is 0.217 and 0.152, respectively. These are not dramatically different in size. But the coefficients in the linear regression for Medu and Fedu are very different, 0.691 and 0.092. How is it that two factors with similar means, variances, and correlations with G3 have such different coefficients?

____

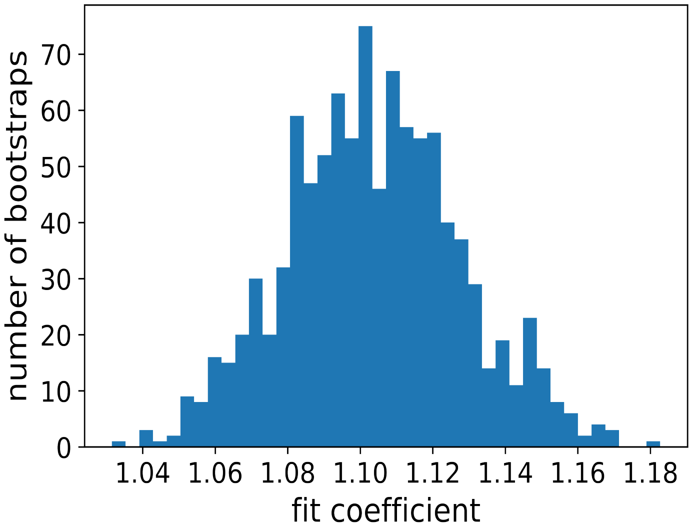
____

____

____

____

1000 bootstrap samples of the database of students were fit to produce give simple least squares coefficients for G3 as a function of one of the other variables in the table.



4d: (3pts) This regression coefficient is not consistent with zero in this population of n=395. This could be the coefficient of G3 with respect to which other variable?

4e: (3pts) How do you make a bootstrap sample?

5a: (3pts) What is the effect of L2 regularization, also called Ridge regression, on multiple-least-squares coefficients?

5b: (4pts) What happens to linear regression fits when the regularization term is "too small" ?