INFO 2950 Fall 2022 Midterm Solution

Instructions

Some students are taking the exam late due to scheduling constraints. Do not discuss the exam unless you are certain that everyone you are talking to has taken it.

You have 70 minutes to complete this exam. Time will be announced and marked on the board. You may use only a writing utensil and paper. If you use any electronic device for any purpose we will immediately confiscate your exam paper. All calculations have been constructed so that you will not need a calculator.

Write answers only in the assigned space on the answer sheet. When you are done, we will collect your answer sheet ONLY. Graders will see the segment of the answer sheet allocated for each problem and nothing else.

Make sure your name and netid are clearly written on every page of the answer sheet, as we will remove staples to scan it.

The answer sheet is intended to provide more than enough space; don’t worry if you don’t fill it. Showing your work may allow us to give you partial credit. Do not spend more than 10 minutes on a problem. If you get stuck, move on and come back later.

Raise your hand if you would like to ask a clarifying question.
Good luck!

$$var(X) = \frac{\sum_i (X_i - \bar{X})^2}{N}$$

$$cov(X, Y) = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

$$corr(X, Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$ (not the same σ!)

$\sigma(t)$ is less than 0.5 when t is negative, and greater than 0.5 when t is positive.

if $Y_i = \alpha + \beta X_i + \epsilon_i$, $\beta = \frac{cov(X,Y)}{var(X)}$

A. Multiple Choice (24 points)

1. Match the following expressions to their output.
   I. float(int("2"))
   II. str(int(2.3))
   III. boolean(int("2.3"))
   IV. int(float("2.3"))

   Possible outputs:
   a. 2
   b. “2”
   c. 2.0
   d. NameError

Answer: I. C, II. B, III. D, IV. A Note that in III, int(“2.3”) gives a ValueError; however, the entire statement gives a NameError since the function passed is boolean() rather than bool().

2. If you delete a file on GitHub and someone simultaneously edits their own copy of that file, what happens after you try to merge branches?
   a. Pull request
   b. Merge conflict
   c. Commit and push
   d. Commit and pull

Answer: **B**

3. For the numpy array given below, what is the value of its `.shape` property?

```
[[[8 0]
  [3 8]
  [9 2]
  [6 3]]
 [[3 6]
  [7 6]
  [5 4]
  [6 2]]
 [[4 9]
  [7 8]
  [2 7]
  [6 7]]]
```

   a. (4, 3, 2)
   b. (3, 4, 2)
   c. (12, 2, 1)
   d. (2, 3, 4)

Answer: B

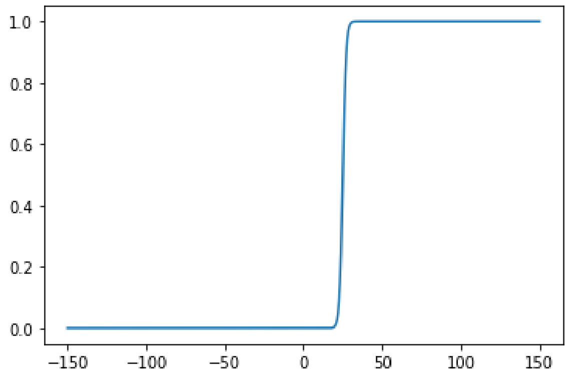
4. Which of these values of a sample has the property that if you convert all the elements less than the value to **-1** and elements greater than the value to **+1**, the sum of all elements of the sample except the value will be **0**?
    a. Mean b. Median c. Mode d. Variance e. None of the above

Answer: B

5. If we remove outliers, which of these necessarily decreases?
    a. Mean
    b. Median
    c. Mode
    d. Variance
    e. None of the above

Answer: D

6. Which of the following is the correct equation for the plot given below? Note: exp(0) = e^0 = 1.0



    a. f(x) = 1 / [1 + exp(-x)]
    b. f(x) = 1 / [1 + exp(-(0.25x))]
    c. f(x) = 1 / [1 + exp(-(x - 25))]
    d. f(x) = 1 / [1 + exp(-(x + 25))]

Answer: C Note that at x=25, we have f(x) = 1 / (1+exp(0)) = 1/2 = 0.5

7. Which of the following is *not* a linear regression?
    a. y ~ ax + b
    b. sqrt(y) ~ α + β x^2
    c. y ~ β x / (1 + β x)
    d. ln(y) ~ α + β log(x)

e. B,C,D are all *not* linear regressions
f. None of the above (i.e., A,B,C, D are all linear regressions)

Answer: C, note that performing a nonlinear transformation on x and/or a nonlinear transformation on y still allows you to run a linear regression on the transformed values of x and y.

8. The matrix given below is a correlation matrix where a, b, c and d are variables:

```
          a         b         c         d
a  1.000000 -0.235053  0.656181  0.595110
b -0.235053  1.000000 -0.583851 -0.471916
c  0.656181 -0.583851  1.000000  0.871202
d  0.595110 -0.471916  0.871202  1.000000
```

Which pair of variables do you suspect is most likely to be collinear?
a. (a, a)   b. (a, b)   c. (b, c)   d. (c, d)

Answer: D, since the magnitude of the absolute value of the pair is highest. (a,a) is not a "pair of variables" since it just shows exact collinearity with itself at correlation 1.0.

B. Describe What's Wrong (12 points)

1. Your coworker asks you to analyze the following "correlation matrix". What are two reasons that make you believe this is not a correlation matrix?

```
      a     b
a   3.2  -4.7
b  -2.3   1.8
```

Answer: The correlation matrix lacks 1's on the diagonals, and the values in the values should be between -1 and 1 (in the matrix, corr(a,a)!=corr(b,b)!=1). Also, the matrix is not symmetric (in the matrix, corr(a,b)!=corr(b,a)).

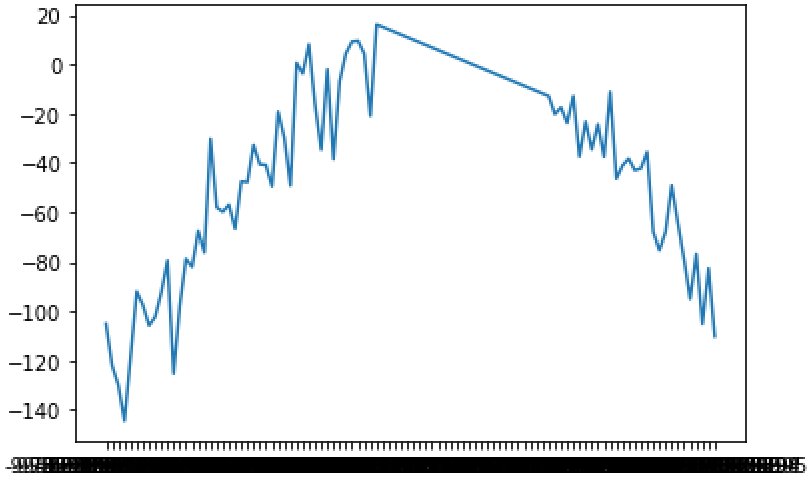
2. What are four things in this dataframe that make you worry?

Fruit	**Is_Organic**	**Cost_USD**	**Popularity_Percent**
Apple	0	1.50	80
Banana	1	1.00	Ninety
Cantaloupe	1	5.99	110
Durian	0	0.000000000001	10
Elderberry	2	7.99	60

Answer:

1. Is_Oraganic == 2 does not take the value of 0 or 1, which is inconsistent with Is_Organic being a binary variable.
2. Cost_USD == 0.00000001 is too small compared to other values and inconsistently represented with a different number of decimal places.
3. Popularity_Percent == Ninety which is a string while the rest of the column is numeric.
4. Popularity_Percent == 110 which is greater than 100 (what one would imagine is the maximum possible value of a "popularity percent").

3. There are a few things wrong with this plot. Pick one, describe what you observe, and make a guess about what might cause it. Propose an idea about how to fix the issue (assume you are correct in your diagnosis).



Answer:

1. If regarding the straight line in the middle of the graph: due to missing values, underlying data needs to be examined to determine why missing values exist (e.g., better values should be imputed).
2. If regarding the "date smear" in the x-axis: due to poor formatting of dates horizontally, they all blend together; it could be fixed by formatting dates to be shorter, at fewer intervals, and/or rotating dates in the x-axis.
3. If regarding lack of clarity in the plot: we can fix it by adding axis labels.

C. SQL (16 points)

You are given the following dataframes **df1** and **df2** regarding the INFO2950 pets, their descriptors, and their corresponding visits and costs of going to the vet.

df1

Pet	**Age**	**Species**
Basil	8	Rabbit
Sparky	2	Cat
Angora	1	Cat

df2

VetVisit	**VetCost**	**Pet**
2021-03-15	45	Basil
2021-08-01	115	Sparky
2022-03-20	55	Basil
2022-07-08	85	Sparky
2022-07-08	20	Angora
2022-07-08	0	Plant

1. Write the SQL statement that will generate the following table:

df3

Pet	**TotalCost**
Basil	100
Sparky	200
Angora	20
Plant	0

Answer: SELECT Pet, sum(VetCost) as TotalCost FROM df2 GROUP BY Pet

2. Write the SQL statement (without using WHERE) that will generate the following table:

df4

Pet	**Age**	**Species**	**TotalCost**
Basil	8	Rabbit	100
Sparky	2	Cat	200
Angora	1	Cat	20

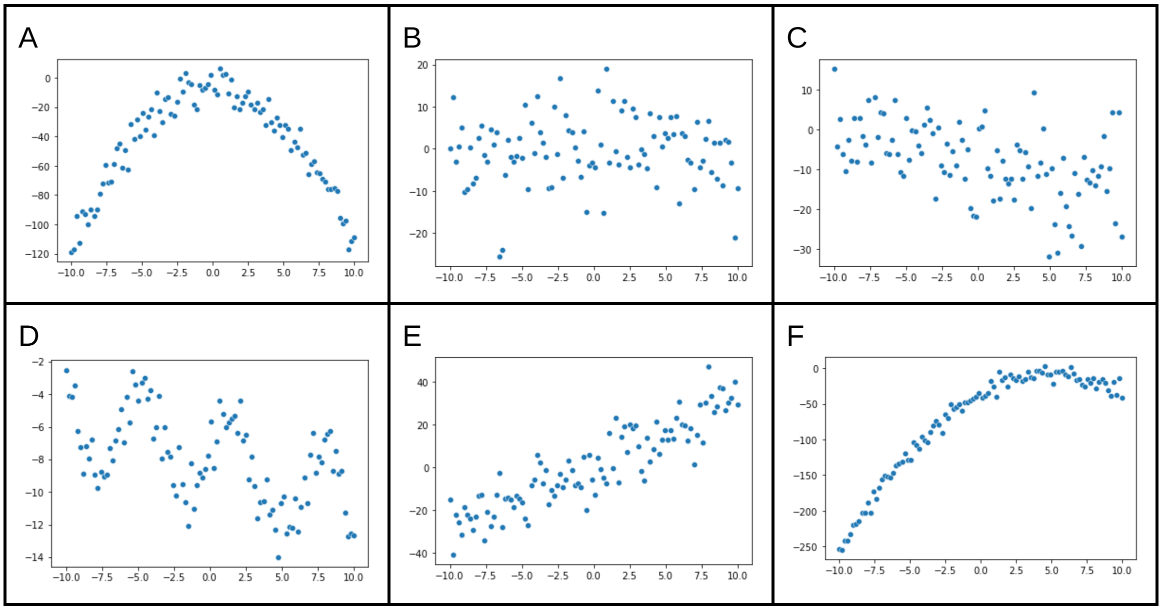
Answer: SELECT df1.Pet, df1.Age, df1.Species, df3.TotalCost FROM df1 LEFT JOIN df3 ON df1.Pet = df3.Pet

3. Write the SQL statement that is similar to **df4** but restricted only to cats, and sorts rows from lowest to highest total cost. Then, tell us the shape of your resulting dataframe.

Answer: SELECT * FROM df4 WHERE Species = “Cat” ORDER BY TotalCost. Recall that the default sort order is ASCENDING (from lowest to highest) already. The shape of the resulting data frame is (2,4).

D. Statistics and Regression (24 points)

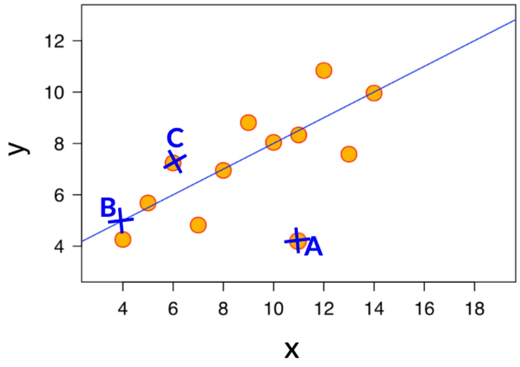
1. For each of the following plots, describe whether the correlation between x and y values is positive, negative, or close to zero.



Answer:

A. Close to zero
B. Close to zero
C. Negative
D. Negative
E. Positive
F. Positive

2. The following figure depicts a scatter plot of (x, y) pairs and a regression line fitted to these data points. There are three points in the plot labeled with A, B and C respectively (in blue). Match the comments listed below (in I, II and III) with points A, B and C:



I. The data point with the largest residual ε
II. The point where α + βx = 5
III. The data point when x = 6

Image source:
https://en.wikipedia.org/wiki/Linear_regression

Answer:
I. A (the distance between the predicted value $\hat{y}$ and the actual value y).
II. B (notice the y coordinate of B equals 5).
III. C (notice the x coordinate of C equals 6).

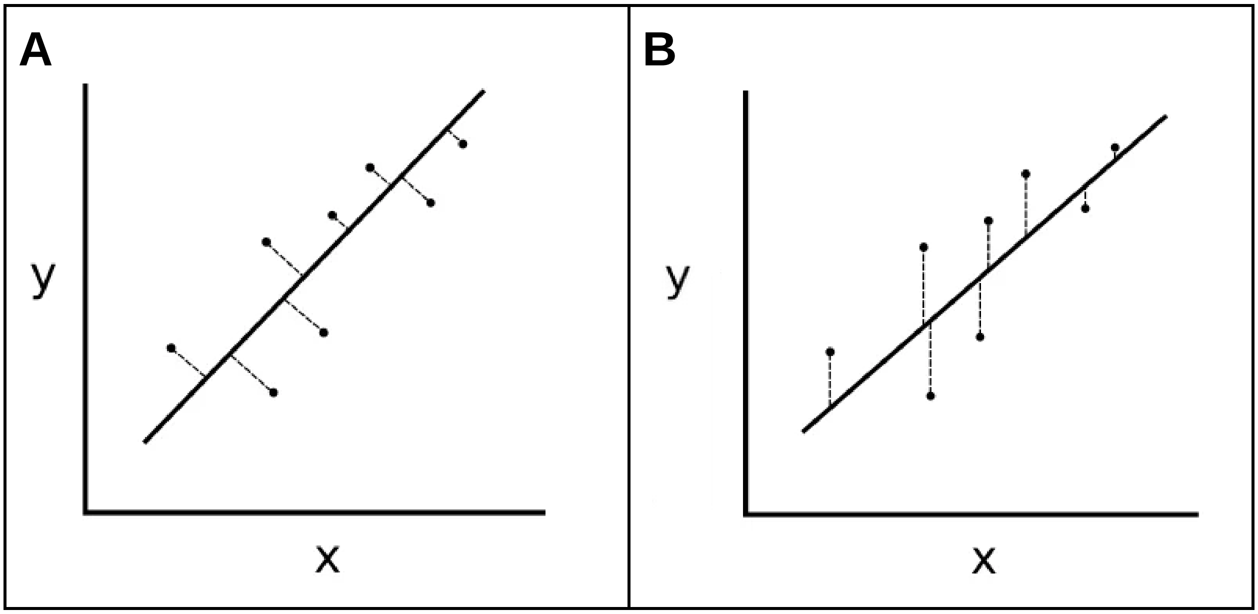
3. The variance of *x* is 2.0, and the variance of *y* is 4.0. The slope β of the regression y ~ x is 1.5. What is the covariance of *x* and *y*? What is the slope of the regression x ~ y? Show your work.

Answer:

1. β for $y \sim x$ is $\frac{cov(x,y)}{var(x)} = 1.5$, where $var(x) = 2$, thus $cov(x, y) = 1.5 * 2 = 3$
2. β for $x \sim y$ is $\frac{cov(y,x)}{var(y)} = \frac{3}{4}$ since cov(y,x) = cov(x,y)

Note: A common error we found in students' answers was claiming the slope of x~y is $\frac{1}{\beta}$. This is incorrect because one cannot simply perform arithmetic on $y = \alpha + \beta * x$ – if switching from y~x to $x \sim y$ you would need to instead minimize on horizontal residuals rather than the usual vertical residuals.

4. Which of the two images below best represents the ordinary least squares regression we learned about in class? Explain why, and specify the mathematical concept that regression minimizes.



(Image source: https://subscription.packtpub.com/book/data/9781786462169/3/ch03lvl1sec32/implementing-deming-regression)

Answer: B, because OLS minimizes the sum of squared error. The dotted lines (indicating distances) should be vertical distances rather than perpendicular to the fitted line.

5. You make a residual plot based on your data.
    a. What do the x and y axes of a residual plot represent?
    b. Describe one characteristic of a residual plot that would indicate to you that your regression does not have problems.
    c. If the characteristic described in (b) is not in your residual plot, how should you improve your regression model?

Answer:
A. The x-axis represents the "fitted value" $\hat{y}$, and the y-axis represents the residual.
B. The points are *randomly* dispersed around the horizontal axis.
C. Perform transformation on existing variables or add additional variables.

6. A logistic regression model has intercept α = 2 and slope β = -3. Is the probability output greater for x = 5 or x = 1? Is the probability for x=1 greater or less than 50? Why?

Answer:
When x = 5, we have $\sigma(-13) = \frac{1}{1+e^{13}}$ while when x = 1, we have $\sigma(-1) = \frac{1}{1+e}$. Both expressions have the same numerator 1, the denominator is greater for $\sigma(-13)$ so the probability is smaller

when x=5 and greater when x=1. When t is negative, the probability σ(t) is less than 0.5. The probability for x=1 is less than 0.5 since t = $\alpha + \beta * x$ = -1, which is negative.

E. Interpreting regressions (24 points)

In the following questions, we provide variables for input x (or multiple input x's) and output y, as well as the regression model that describes the relationship between those variables. For each of these questions, interpret using the relevant interpretation strategies:

a. *Summarize the relationship between variables (specify units and reference variables; an exponentiated lookup table for different numbers is provided at the end of this section)*
b. *Predict y-hat (try plugging in values of x that allow you to simplify the resulting y-hat calculations; you may leave predictions as unreduced numeric expressions for this exam)*
c. *Describe outliers and oddities*

1. y = 4.5 + 2.7x_1 - 4.2x_2

    y = daily profit from sale of umbrellas ($ USD)
    x_1 = daily average rainfall (millimeters of rain)
    x_2 = weekend binary (x_2 = 1 if weekend, x_2 = 0 if weekday)

Answer:
A. Summarize: For an additional millimeter of daily average rainfall, we expect a $2.70 increase in daily profit from the sale of umbrellas, all else equal. For weekend days (as opposed to weekdays), we expect to see a $4.20 decrease in daily profit from the sale of umbrellas, all else equal.
B. Predict: For some value x_1 of rainfall on a weekend, we predict that daily profit from the sale of umbrellas is $\$4.50 + \$2.70 * x_1 - \$4.20$. For some value x_1 of rainfall on a weekday, we predict that daily profit from the sale of umbrellas is $\$4.50 + \$2.70 * x_1$.
C. Oddities: We can never predict a profit of less than $0.30 for umbrellas on weekends, which seems unreasonable (e.g., what if sometimes people never buy umbrellas?). Additionally, this model would imply that as the amount of rainfall goes up, the profit of umbrella sales do too, but we might expect that during hurricanes (high x_1 value) people will stay inside and not buy umbrellas (which would imply a lower $\hat{y}$ instead of higher).

2. ln(y) = -2.05 + 1.95x_1

    y = number of BeReal app users (Approximation from data loosely based on https://www.onlineoptimism.com/blog/bereal-stats-app-figures-data-be-real-numbers-to-know/)
    x_1 = quarters (from Q1-2020 to Q2-2022), where each quarter represents 3 months

Answer:
A. Summarize: From one quarter to the next, we expect that the number of BeReal users will be multiplied by $e^{1.95} \approx 7.03$. Or, a 1 unit change in quarter is associated with a $100 * (7.03 - 1)\% = 603\%$ change in BeReal users.
B. Predict: For the 0th quarter (Q1-2020) we predict that BeReal had e^{a} = $e^{-2.05}$ = 0.13 users. For the nth quarter, we predict that BeReal had $e^{1.95*n - 2.05}$ users.
C. Oddities: This model should predict integer counts of users but instead predicts numbers with decimals. This model predicts that going forwards in time (after 2022) can give you exponentially many expected BeReal app users (e.g., more than the population of the world, which is unlikely). Note: It's not true that going backwards in time can give you negative users, since the exponentiation of a negative number still gives you a positive number.

3. y = σ(4.71 - 0.72x$_1$)

    y = whether a college student is not a senior (binary, where y = 1 if freshman/sophomore/junior, and y=0 if senior)
    x$_1$ = number of jobs the student applied to (numeric)

Answer:
A. Summarize: For each additional job you apply for, we expect the odds of you being not-a-senior to be multiplied by $e^{\beta} = 0.49$. I.e., a 1 unit increase in the number of jobs applied to is associated with a $100 * (e^{\beta} - 1) = 51\%$ decrease in the probability that you're not a senior.
B. Predict: The probability that you're not a senior, given that you've applied to 0 jobs, is $(\frac{e^{\alpha}}{e^{\alpha}+1}) = 0.99.$
C. Oddities: Apply your domain knowledge – maybe you think that non-seniors often apply to jobs like internships, so the intercept seems fishy. Or, maybe the construction of y as a binary variable is not meaningful since both juniors and seniors apply for a lot of jobs. There is no constraint on what value x$_1$ can take, so we want to be careful about making sure we don't input negative numbers for x$_1$ or massive values of x$_1$ that exceed the number of jobs in the world.

4. y = 6.0 + 54.0x$_1$ + 41.5x$_2$- 6.0x$_1$*x$_2$

    y = customer satisfaction of skincare product (numeric from 0 to 100)
    x$_1$ = whether Vitamin C is in the skincare product (binary)
    x$_2$ = whether Vitamin E is in the skincare product (binary)

Answer:
A. Summarize: Per Lecture 11, there should be no summary for regressions with interactions; it is more useful to predict than summarize in this case. Note: It's not correct if you say having Vitamin C and E together results in a decrease in customer satisfaction.
B. Predict: When there is no vitamin C and no vitamin E in the product, we predict customer satisfaction is lowest, at $6.0 + 54.0 * 0 + 41.5 * 0 - 6.0 * 0 = 6$. When there is only vitamin C in the product, we predict customer satisfaction is$6.0 + 54.0 * 1 + 41.5 * 0 - 6.0 * 0$ = 60.

When there is only vitamin E in the product, we predict customer satisfaction is $6.0 + 54.0 * 0 + 41.5 * 1 - 6.0 * 0 = 47.5$. When there is both vitamin C and vitamin E in the product, we predict customer satisfaction to be the highest, at $6.0 + 54.0 * 1 + 41.5 * 1 - 6.0 * 1 = 95.5$.
C. Oddities: There are only four possible values that the $\hat{y}$ outcome can take, which might not be realistic. For example, there is no scenario where customer satisfaction is anything other than 6, 47.5, 60, or 95.5. Also, maybe Vitamin C and E are better modeled as numeric values (e.g. the amount in a product) rather than binaries.

Exponential lookup table (you may use as many or as few of these as needed):

n	**e^n**	**σ(n)**
-6.0	0.002	0.002
-4.2	0.01	0.01
-2.05	0.13	0.12
-0.72	0.49	0.33
1.95	7.03	0.88
2.7	14.88	0.94
4.5	90.02	0.99
4.71	111.05	0.99
6.0	403.4	1.00