INFO 2950 Fall 2023 Midterm Solutions

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 (including a calculator) 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. ONLY your answer sheet, and not your question sheet, will be graded. The exam will be graded out of 100 total points.

Make sure your name and netid are clearly written on every page of the answer sheet, as we will remove staples to scan it. If you do not write your answers clearly, they will not be scanned well and may be graded incorrectly.

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

I. Programming with Data: Multiple Choice (8 points)

1. *(2 points)* What will be the following expression's output?

   `type("29.50")`

   A. An error (**`.dtype`** should be used)
   B. **`float`**
   C. **`string`**
   D. **`NaN`**

Answer: C. *Slides: Lec 1 slides 27-40, Lec 2 slides 82-86*

2. *(2 points)* Which of the following is not necessary when analyzing meaningful time series data?

   A. Having regularly spaced and chronological time labels
   B. Having corresponding data per time step
   C. Converting the time label to a string
   D. Having unique time labels
   E. Dealing with missing values

Answers: C. *Slides: Lec 6 slides 9, 28*

3. *(2 points)* Which of the following would convert the **`Date`** column to a datetime object, and replace the original column with the new series?

   A. **`df[Date] = pd.to_datetime(df["Date"])`**
   B. **`df = pd.to_datetime(df["Date"])`**
   C. **`df["Date"] = pd.to_datetime(df["Date"])`**
   D. **`df["Date_New"] = pd.to_datetime(df["Date"])`**

Answer: C. *Slides: Lec 6 slides 29, 48*

4. *(2 points)* Both correlation and covariance are indicators of X and Y relationship directions that are between -1 and 1. True or False?

Answer: FALSE. *Slides: Lec 6 slide 75*

II. Programming with Data: Short Answer (20 points)

1. *(2 points)* Given **`arr = np.array([[3,2,1], [6,5,4], [9,8,7]])`**, how would you index for the value **`4`** using numpy? Fill in the blank: **`arr[___]`**
Answer: arr[1,2] - indexing in Python starts at 0. *Slides: Lec 2 slides 32-34*

2. *(2 points)* What is the output of df.shape?

```
    df = pd.DataFrame({"a": [1,2], "b": [3,4], "c": [5,6]})
```

Answer: (2, 3). *Slides: Lec 2 slide 49, Lec 3 slide 75*

3. *(4 points)* In your answer sheet, circle the four cells that would require cleaning in the following dataframe.

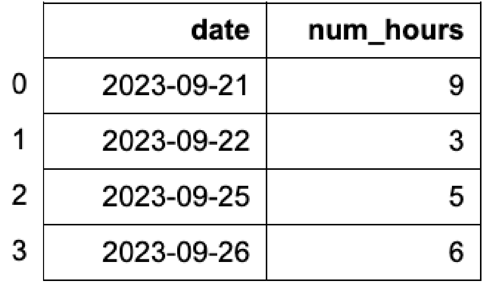
Name	**Major**	**Graduation Year**	**Credits Taken**	**Graduated Already?**
Jules		2022	85	1
Kiki	Information Science	TwentyTwentyFour	105	0
Jon	Information Science	2021	-5	1
Leo	Animal Science	2025	15	0
Bella	Philosophy	2026	30	0
Toffi	Food Science	2026	45	2

Answer:
1. GraduationYear, "TwentyTwentyFour" --> string not a year
2. Credits Taken, -5 --> what do negative credits mean?
3. Graduated Already, 2 --> binary column with a 2 in it
4. Major, "" --> missing data

Slides: Lec 1 slide 81, Lec 6 slide 42

4. *(4 points)* Given the dataframe `office_hours`, which graph would be the output when we execute the following code: `office_hours.plot("date", "num_hours")`? Circle whether graph A or B would be output, and explain why in a few words.

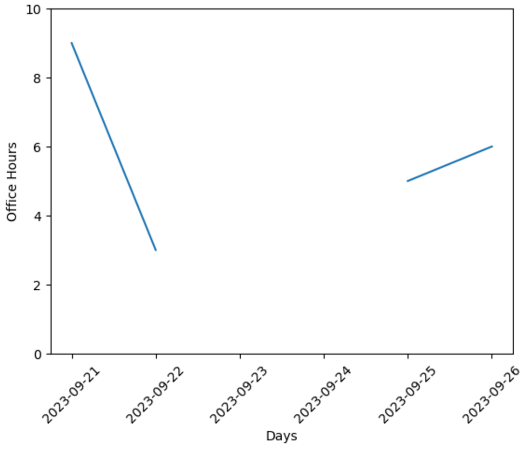
`office_hours:`

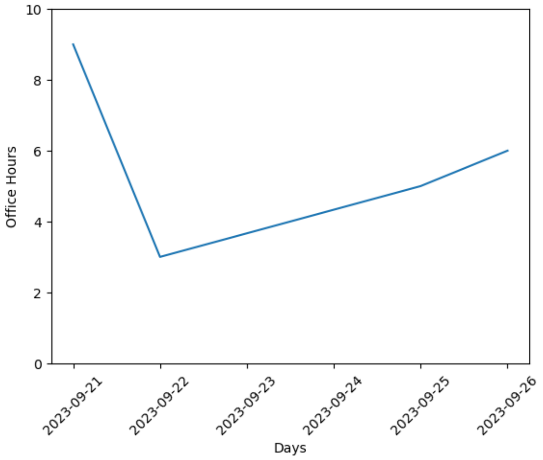


	date	num_hours
0	2023-09-21	9
1	2023-09-22	3
2	2023-09-25	5
3	2023-09-26	6

A.

B.





Answer: Graph B: days are missing but we do not have NaN values.
Slides: Lec 6 slides 59, 68

5. *(8 points)* There are four syntax errors in the following program involving a dataframe with columns `temp_value` and `temp_class`. Write the line number and the reason for each error. There may be multiple errors or no errors in each line.

```
import pandas as pd
import duckdb

df = pandas.read_csv(dataframe.csv) # LINE 1
print(df[temp_value].median()) # LINE 2
print(df["temp_value"].mean()) # LINE 3
high_df = duckdb.sql("SELECT * WHERE temp_class = 'high'").df() # LINE 4
```

Answer

- Error 1: Line 1, no quotes *Slides: Lec 2 slide 63*
- Error 2: Line 1, pandas instead of pd *Slides: Lec 2 slide 43*
- Error 3: Line 2, no quotes *Slides: Lec 6 slide 48*
- Error 4: Line 4, not loading from a data frame (no FROM statement) *Slides: Lec 2 slides 102-105*

III. Programming with Data: SQL (15 points)

You are given the following dataframes **df1** and **df2** regarding the INFO2950 pets, their descriptors, and costs from different purchases at the pet store.

df1

Pet	**Age**	**Species**

df2

PetStoreVisit	**PetStoreCost**	**PetStoreCategory**	**Pet**

Libby	7	Cat
Juno	2	Cat
Pluto	NaN	Plant

2023-03-15	30	Food	Libby
2023-03-15	20	Litter	Libby
2023-03-20	10	Toy	Juno
2023-07-08	20	Food	Juno
2023-07-15	20	Litter	Juno
2023-07-15	10	Toy	Libby

1. *(5 points)* Write the SQL statement that fills in the blank to generate **df3** (using **df2** and/or **df1**).

```
df3 = duckdb.sql("________________________________").df()
```

df3

Pet	**TotalCost**
Libby	60
Juno	50

Answer: SELECT Pet, SUM(PetStoreCost) AS TotalCost FROM df2 GROUP BY Pet
Slides: Lec 4 slides 49-54

2. *(5 points)* Write the SQL statement that fills in the blank to generate **df4** (using **df3,** and either **df1** or **df2**).

```
df4 = duckdb.sql("________________________________").df()
```

df4

Pet	Age	Species	TotalCost
Libby	7	Cat	60
Juno	2	Cat	50

Answers:
SELECT df1.Pet, Age, Species, TotalCost FROM df3 INNER JOIN df1 ON df3.Pet = df1.Pet

Other equivalent answers work too, e.g. selecting df3.Pet instead of df1.Pet; specifying df1.Age, df1.Species, or df3.TotalCost are all fine. You could also flip the order of the ON statement (df1.Pet = df3.Pet), or assign df3 and df1 to new names, e.g.

SELECT a.Pet, Age, Species, TotalCost FROM df1 AS a INNER JOIN df3 AS b ON a.Pet = b.Pet

You could also do a left join onto df3 (but not a left join onto df1). The correct left join would be:
SELECT a.Pet, Age, Species, TotalCost FROM df3 AS a LEFT JOIN df1 AS b ON a.Pet = b.Pet

Slides: Lec 5 slides 132-140

3. *(5 points)* Fill in the table with the headings and data that would result from the following SQL expression.

```
df5 = duckdb.sql("SELECT PetStoreCategory, AVG(PetStoreCost) AS AverageCost
FROM df2 GROUP BY PetStoreCategory ORDER BY AverageCost DESC").df()
```

df5

Answer:
Slides: Lec 4 slides 57-59
df5

PetStoreCategory	**AverageCost**
Food	25
Litter	20
Toy	10

IV. Regression: Multiple Choice (12 points)

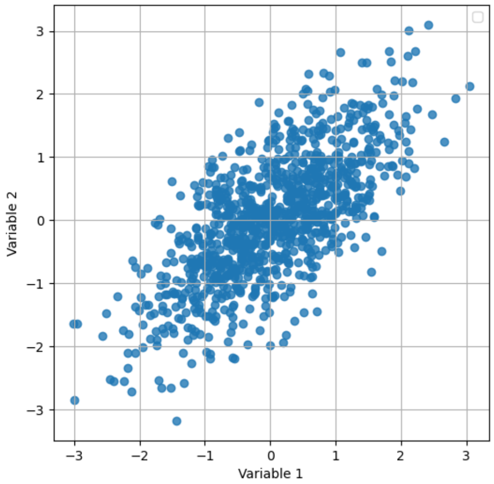
1. *(2 points)* What are possible ways of finding α and β in a regression?
   A. Doing it by hand (with calculus) for a linear regression
   B. Doing it by hand (with calculus) for a logistic regression
   C. Using Python packages
   D. A and C

E. B and C
F. All of the above

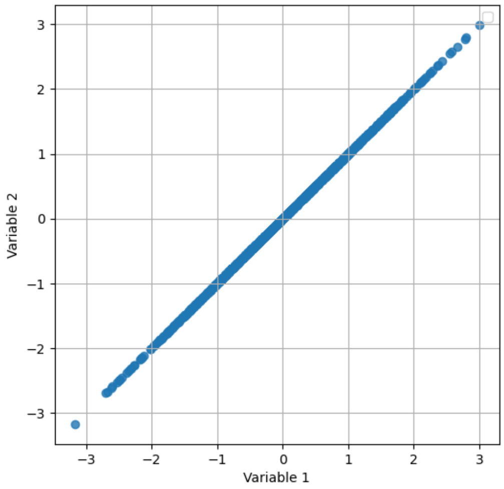
Answer: D. *Slides: Lec 7 slides 50-77, 99, Lec 9 slides 114-116*

2. *(2 points)* If you rank the following plots in order from smallest to largest covariance, which of the following is the correct order?

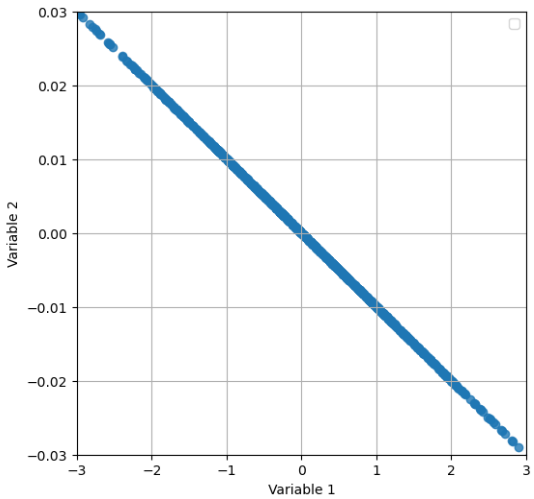
i.



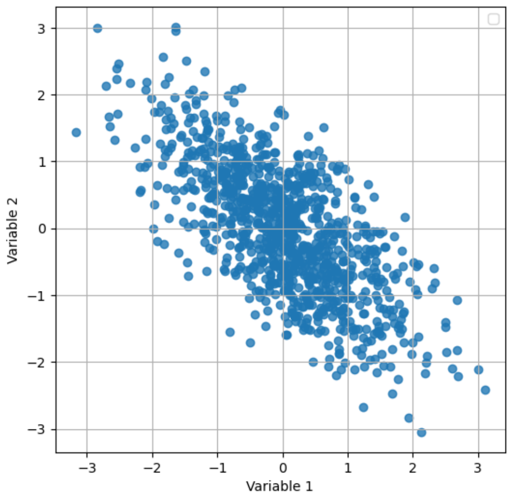
ii.



iii.



iv.



A. iii, iv, i, ii
B. ii, i, iii, iv
C. iv, iii, i, ii
D. ii, i, iv, iii
E. More information is needed to answer this.

Answer: C. *Slides: Lec 5 slides 59-60, 68-70*

3. *(2 points)* Which term represents the “least squares” component that one tries to minimize when using ordinary least squares (OLS) regression?

A. $\sum\left(x_i - \hat{x}_i\right)^2$ **B.** $\sum | y_i - \hat{y}_i |$ **C.** $\sum \hat{\varepsilon}_i$

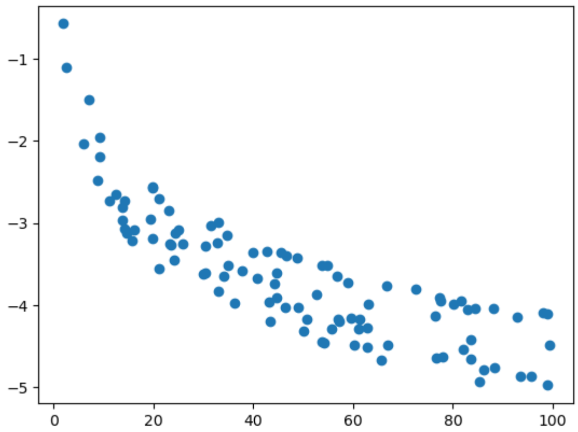
D. $\sum\left(y_i - \hat{y}_i\right)^2$ **E.** $\frac{cov(y,\ x)}{var(x)}$

Answer: D. *Slides: Lec 7 slides 61-64*

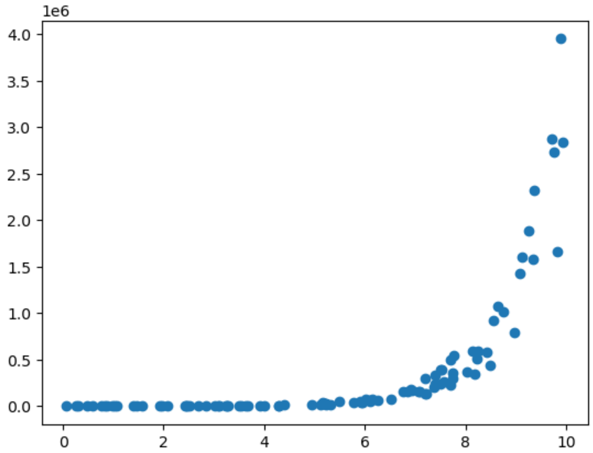
4. *(3 points)* Each of the following plots would be best modeled by one of the three following equations. Each equation was used once. Select the answer that matches the plots to the equations correctly.

1) $y_i=\alpha+\beta x_i+\epsilon_i$ **2)** $y_i=\alpha+\beta\log(x_i)+\epsilon_i$ **3)** $\log(y_i)=\alpha+\beta x_i+\epsilon_i$

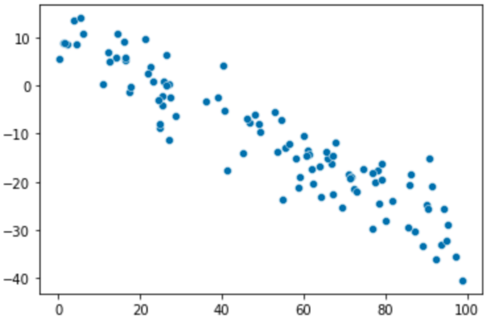
i.



ii.



iii.



Options:

A. 1=iii, 2=ii, 3=i

B. 1=iii, 2=i, 3=ii

C. 1=ii, 2=iii, 3=i

D. 1=i, 2=iii, 3=ii

Answer:

Option B. *Slides: Lec 8, slide 119; Lec 11, slides 51-55.*

5. *(3 points)* Which of the following components should you include in your interpretation of y ~ x_1 + x_2 + x_3 when summarizing the relationship between variables? y represents the cost of a product, and x_1 represents the weight of the product. The product can be one of three items: rock, paper, and scissors. The dummy variables are as follows: x_2 represents being a rock product, and x_3 represents being a paper product.

A. “All else equal,...”
B. “we expect y to increase/decrease by β_2 if it is a rock product relative to being a scissor product.”
C. “we expect y to increase/decrease by β_2 if it is a rock product relative to being not a rock product.”
D. A and B
E. A and C

Answer: D. *Slides: Lec 10 slide 95 and 122-127, Lec 11 slides 69-90.*

V. Regression: Short Answer (29 points)

1. *(2 points)* Fill in the blanks for a linear model that represents *Gimme! Coffee*’s hot chocolate sales. There are only two types of days that can be represented by x: either it snows (x=1) or it does not snow (x=0). If it does not snow, the model predicts that *Gimme! Coffee* will sell 10 hot chocolates. If it does snow, the model predicts that *Gimme! Coffee* will sell 35 hot chocolates.

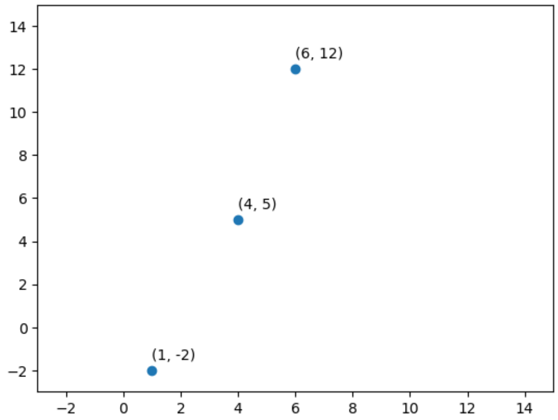
y = ___ + ___ x

Answer:

Equation: y = 10 + 25x
When it snows (x = 1): 35 hot chocolates. When it doesn’t snow (x=0): 10 hot chocolates.
Slides: Lec 7 slides 25-29

2. Using the following two linear models and three data points:

Model 1: y = 2x - 3
Model 2: y = 3x - 6



i. *(4 points)* Fill in the table in your answer sheet by computing ŷ and epsilon for each x_i.

ii. *(2 points)* Compute the mean squared error for each model. You may leave your answer in unreduced fraction form.

iii. *(2 points)* Which model fits the data better? (Circle your answer).

Model 1: y = 2x - 3 **Model 2: y = 3x - 6**

x	**y**	**ŷ (model 1)**	**ε (model 1)**	**ŷ (model 2)**	**ε (model 2)**
1	-2				
4	5				
6	12				

Answers:
i.

x	**y**	**ŷ (model 1)**	**ε (model 1)**	**ŷ (model 2)**	**ε (model 2)**
1	-2	-1	-1	-3	1
4	5	5	0	6	-1
6	12	9	3	12	0

ii. Model 1 Mean Squared Error: $(-1^2 + 0^2 + 3^2)/3 = 10/3$
Model 2 Mean Squared Error: $(1^2 + -1^2 + 0^2)/3 = 2/3$

iii. Model 2 is better fit.
Slides: Lec 6 slides 125-130, Lec 7 slides 60

3. For some binary outcome Y:

i. *(4 points)* Fill in the missing values in the table. Express probabilities as fractions.

Row #	Log odds	Probability	Odds
1			**1:1**
2	**-0.30**		**1:2**
3	**2.94**		**19:1**

ii. *(1 point)* Which Row # has the highest Pr(Y=1)/Pr(Y=0)?

Answer: *Slides: Lec 10 slides 9-10, 15-25*

Log odds	Probability	Odds
0	**0.5**	1:1
-0.30	⅓	1:2
2.94	**0.95**	19:1

Row #3.

4. *(8 points)* Fill in the blanks to derive how a change in x would affect the output *y* for a linear-log model.

Step 1: y = a+b*ln(x)

Step 2: Define new variable $\mathbf{x_{new}}$**=** ________

Step 3: Define new variable y_{new} so that y_{new} = a+b*ln(x_{new})

Step 4: Rewrite the right-hand side to be in terms of x instead of x_{new}

$\mathbf{y_{new}}$ **=** ______________

Step 5: Rewrite the right-hand side to be in terms of y instead of x

$\mathbf{y_{new}}$ **=** ______________

Step 6: Write the change between y_{new} and y, and **explain in words** how you would interpret this: ______________

Answer:

EITHER:

x_{new}= x*e

y_{new}= a+b*log(x*e)=a+b*log(x) +b*log(e)=a+b*log(x)+b

y_{new}= y+b

y_{new}-y = b. Multiplying x by e yields a b unit change in y.

OR:

x_{new}= x*1.01

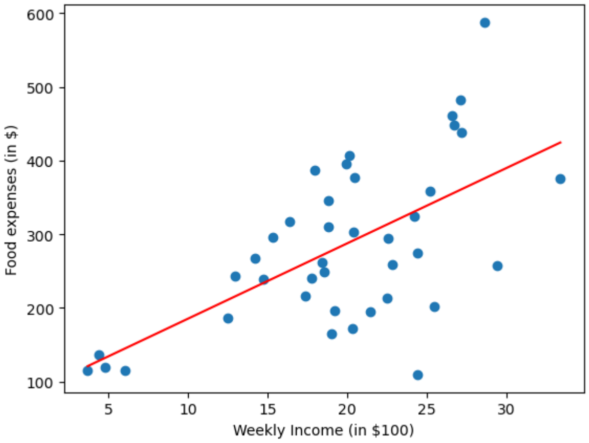
y_{new}= a+b*log(x*1.01)=a+b*log(x) +b*log(1.01)≈a+b*log(x)+b*0.01

y_{new}= y+0.01*b

y_{new}-y = 0.01b. Increasing x multiplicatively by 1% yields a 0.01*b unit change in y.

Slides: Lec 9 slides 10-19 OR Lec 9 slides 21-26, Lec 11 slides 59-67

5. The below figure shows data points and a regression line using weekly income to predict food expenditure.

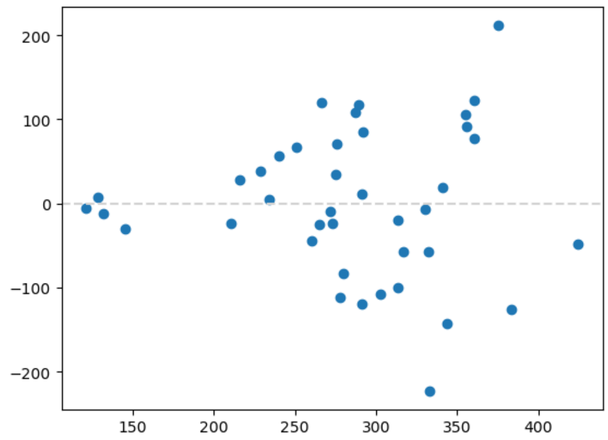


i. *(2 points)* Is Figure A or B the corresponding residual plot?

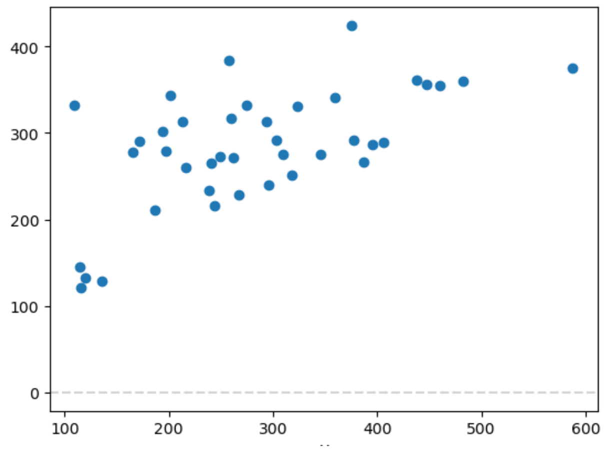
ii. *(2 points)* What do the X and Y axes of a residual plot represent? (Answer in 2 sentences or fewer.)

iii. *(2 points)* Would you trust a regression fit to the data shown in this problem? Why or why not? If not, how would you solve this issue? (Answer in 2 sentences or fewer.)

A:



B:



Answer: *Slides: Lec 8 slides 86-97, 108-115*

i. A

ii. The x axis represents the $\hat{y}$ value predicted by the model. The y axis represents the residual, which is the difference between the $\hat{y}$ (predicted) value and the actual (fitted) value.

iii. No, I would not trust the regression model using the raw data because the residual plot exhibits a "fan" shape common to heteroskedastic data (the variance of the residuals increase over the x-axis). I would solve this with a transformation (e.g., taking the logarithm or either / both axes.

VI. Regression: Interpretations (16 points)

An exponential lookup table is provided on the following page.

1. Suppose we have the following regression model:

x = number of days since the start of October (i.e. October 1st = day 0, etc.)
y = quantity of Halloween candy sold
$\log_e(y) = x + 2$

Interpret the relationship between days since the start of October and the amount of Halloween candy sold. Make sure to describe using full sentences.

i. *(3 points)* Summarize the relationship between variables.
ii. *(3 points)* Make a prediction for x = 0, and for x = 1.
iii. *(2 points)* Inspect oddities.

Answers:

i. For each additional 1 day that goes by, we expect the quantity of Halloween candy sold to be multiplied by $e^{\beta}=e^{1}\approx 2.72$ (i.e., a 172% increase in Halloween candy sold).

ii. On the first day of October (day x=0), our model predicts that $e^{\alpha}=e^{2}\approx7.39$ Halloween candies will be sold. On the second day of October (day x=1), our model predicts that $e^{\alpha+1}=e^{3}\approx20.1$ Halloween candies will be sold.

iii. The model predicts e^{x+2} candy will be sold on day x, meaning that anytime before 3 days before the beginning of October, the model predicts less than a single piece of candy to be sold. Domain knowledge states that the days after Halloween would see an immediate decline in candy sales, but the model would predict the amount to continue increasing at an extremely high rate (the day after Halloween, the model predicts $e^{34}\approx5.8e14$ candy sales). Note: this model will never predict a negative y; any answers claiming this is the case should lose points.

Slides: Lec 9 slides 5-9

2. You build a logistic regression model to predict whether or not it is fall in Ithaca based on the percentage of colorful (non-green) leaves on trees.

x = percentage of colorful leaves *(x = 70 means 70% of leaves are fall colored)*
y = 1 if it is fall; 0 if any other season
y ~ σ(-1 + 3x)

Interpret the relationship between percentage of colored leaves and the current season being fall. Make sure to describe using full sentences.

i. *(3 points)* Summarize the relationship between variables.

Note: You should only pick one of these phrases to use in your summary:

"If [x] increases by 1 percent…" means going from, e.g., x=50 → x=50.5
"If [x] increases by 1 percentage point…" means, e.g., x=50 → x=51

ii. *(3 points)* Predict the probability of fall when x = 0.
iii. *(2 points)* Inspect oddities.

Answer:

i. For a one unit increase in the percentage of colorful leaves (i.e., an increase of 1 percentage point, not an increase of 1%), we expect the odds of it being fall (i.e., the probability of it being fall, divided by the probability of it not being fall) to be multiplied by $e^{\beta}=e^{3}\approx20$.

ii. When the percentage of colored leaves is 0, the model predicts that the probability that it is fall is $e^{\alpha}/(e^{\alpha}+1) = e^{-1}/(e^{-1}+1)\approx0.27$.

iii. The model allows us to input negative percentages of leaves, or > 100% of leaves, which is nonsensical. It might seem a little fishy that even with 0 leaves our model still predicts a relatively high (27%) chance of it being fall. Additionally, our model might benefit from other input variables such as location or types of trees, since some areas may have coniferous trees that never change colors. Note: this model will never predict a value of *y* that's *not* between 0 and 1; any answers claiming this is the case should lose points.

Slides: Lec 9 slides 123-126, Lec 10 slides 30-32.

Extra Credit

(2 points) What is the correct way to tag your homework on Gradescope?

A. Just the problem description.
B. The problem description, the code, and the output.
C. Just the code and the output.

Answer: C. *Slides: Lec 7 slide 3.*

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

n	**e^n**	**σ(n)**
-3	0.05	0.05
-2	0.14	0.12
-1	0.37	0.27
1	2.72	0.73
2	7.39	0.88
3	20.1	0.95
31	3e13	1.00
32	8e13	1.00
33	2e14	1.00
34	5e14	1.00