

**Question 1: EDA warmup (7 Points)**

A.(5) 200 adults shopping at a supermarket were asked about the highest level of education they had completed and whether or not they smoke cigarettes. Results are summarized in the table.

	Smoker	Non-smoker	Total
High school	32	61	93
2 yr college	5	17	22
4+ yr college	13	72	85
Total	50	150	200

i.(1) What percent of the shoppers were smokers with only high school educations?

What percent of the shoppers with only high school educations were smokers?

What percent of the smokers had only high school educations?

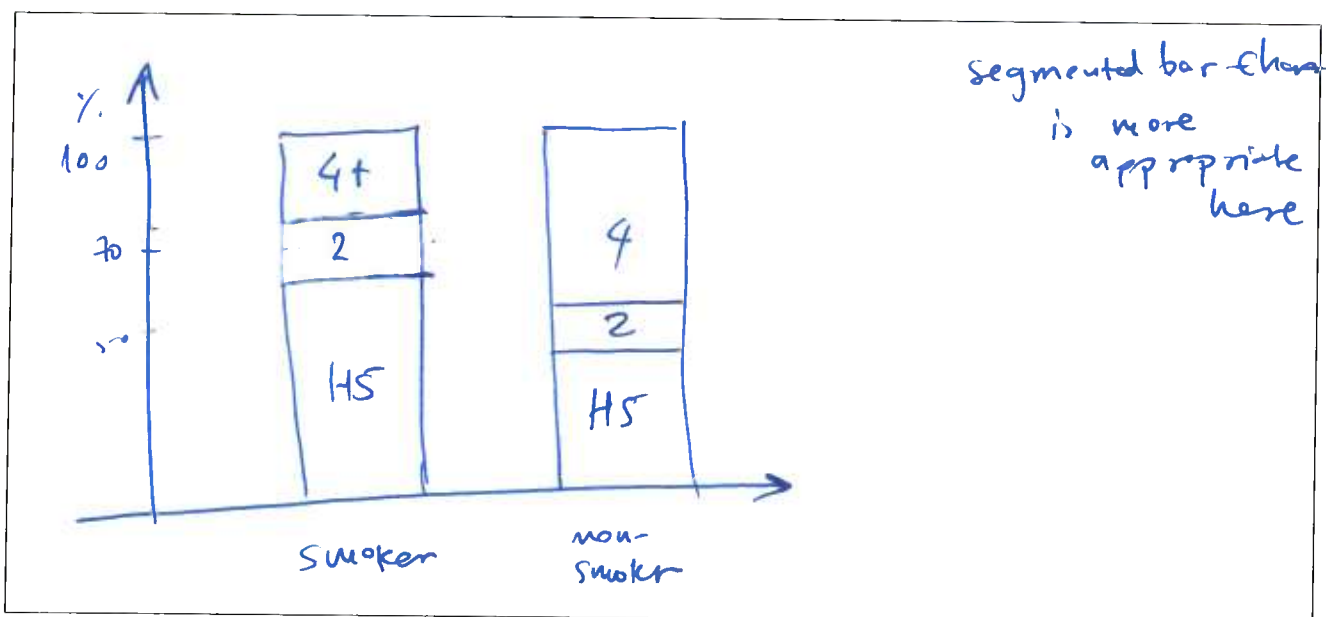
$$(a) \quad 32/200 = 16\%$$

$$(b) \quad 32/93 \approx 34.4\%$$

$$(c) \quad 32/50 = 64\%$$

ii.(2) Sketch an appropriate graph/plot to for comparing education level among smokers and non-smokers.

Label your graph clearly.



Student name:

Student ID:

Page 3 of 15

Data Analysis 2017/2018

iii.(2) Do these data suggest there is an association between smoking and education level? Give *statistical evidence* to support your conclusion. Does this indicate that students who start smoking while in high school tend to give up the habit if they complete college? Explain *shortly*.

(a) There is possible association between smoking & education level

64% of smokers had only a HS diploma

40.7% of non-smokers had only a HS diploma

Similarly, 26% of smokers had 4+ years college compared to 48% of non-smokers

(b) No, this data were collected at one time, about 2 different groups (smokers/non-smokers). We have no idea if smoking behavior changes over time

B.(2) An automobile service shop reported the summary statistics shown for repair bills (in €) for their customers last month.

i.(1) Were any of the bills outliers? Show (using statistical evidence) how you made your decision.

Min	27
Q1	88
Median	132
Q3	308
Max	1442
Mean	284
SD	140

Yes.  $IQR = Q3 - Q1 = 308 - 88 = 220$

Upper Fence =  $Q3 + 1.5 IQR = 638 < 1442$

which means that there are (at least one) outliers

ii.(1) After checking out a problem with your car the service manager gives you an estimate of "only €90." Is the manager right to imply that your bill will be unusually low? Explain briefly.

No. €90 is higher than Q1 (i.e. 25% of all bills)

so it's not unusually low

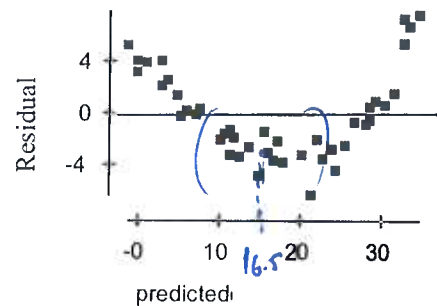
**Question 2: Let's (re)discover penicillin (9 Points)**

Doctors studying how the human body assimilates medication inject some patients with penicillin, and then monitor the concentration of the drug (in units/cc) in the patients' blood for seven hours. First, they tried to fit a linear model. The regression analysis and residuals plot are shown below.

Dependent variable is: **Concentration**  
 No Selector  
 R squared = 90.8% R squared (adjusted) = 90.6%  
 s = 3.472 with 43 - 2 = 41 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	4900.55	1	4900.55	407
Residual	494.199	41	12.0536	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	40.3266	1.295	31.1	$\leq 0.0001$
Time	-5.95956	0.2956	-20.2	$\leq 0.0001$



A.(2) Explain what the value of R squared (90.8%) means in this context and what is the difference from R squared adjusted.

90.8% in the variance of penicillin concentration can be explained by "time" variable

B.(1) Using this model, estimate what the concentration of penicillin will be after 4 hours.

$$\hat{y} = 40.3266 - 5.95956 \times 4 \approx 16.5 \text{ units/cc}$$

C.(2) Is that estimate likely to be accurate, too low or too high? Use the residual plot to explain.

See figure: Predicted value (4 hour) is in an area of negative residuals meaning that the estimate will be too high

Student name:

Student ID:

Page 5 of 15

Data Analysis 2017/2018

Now the researchers try a new model, using the re-expression  $\log(\text{Concentration})$ . Examine the regression analysis and the residuals plot below.

Dependent variable is: **LogCnn**

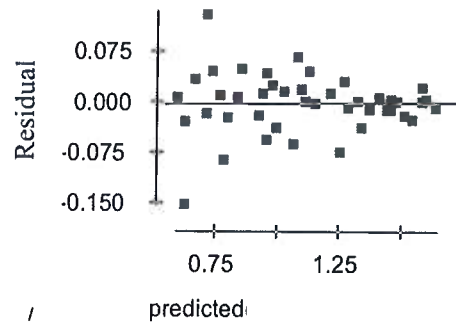
No Selector

R squared = 98.0% R squared (adjusted) = 98.0%

s = 0.0451 with 43 - 2 = 41 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	4.11395	1	4.11395	2022
Residual	0.083412	41	0.002034	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	1.80184	0.0168	107	$\leq 0.0001$
Time	-0.172672	0.0038	-45.0	$\leq 0.0001$



D.(1) Is this model better than the previous one? Explain why you think yes or no.

Yes  
(a) residuals show no pattern (random)  
(b) better adj.  $R^2$

E.(1) Using this new model, estimate the concentration of penicillin after 4 hours.

$$\log_{10} C = 1.80184 - 0.172672 \times 4 = 1.11152$$

$\Downarrow$

$$C = 10^{1.11152} \approx 12.9 \text{ units/cc}$$

F.(2) Explain (in the context of this specific problem) what the coefficient of "time" means.

Every 1 hour passes  $\Rightarrow$  the logarithm of concentration is reduced by  $-0.172672$

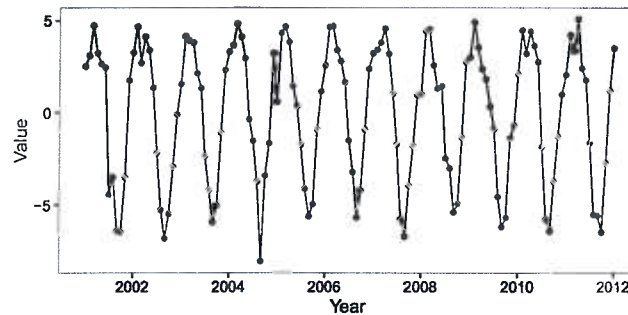
1. vi - A :  $(x_1, x_2) \Leftrightarrow (ax_1, x_2)$  : shrinks in the  $x_1$  direction  
 FT:  $\tilde{f}(f(ax_1, x_2)) = \frac{1}{a} \tilde{f}(f(\frac{x_1}{a}, x_2))$  : stretched in  $x_1$  direction and the whole figure stretches
2. ii - F :  $f(x_1, ax_2)$  similar with 1.
3. v - C : shift  $x_1$  by  $a$  to the left  
 FT: only a phase change  $\Leftrightarrow$  same
4. iv - D : A: rotation by  $\pi/6 \Rightarrow$  similar with FT
5. i - E : similar with 4.
6. iii - B : well, only one choice left ☺  
 $\tilde{f}(f(x_1, x_2) \cdot \text{sinc}(ax_1) \cdot \text{sinc}(ax_2)) = \tilde{f}(f(f_1, f_2) \cdot \Pi_a(f_1) \Pi_a(f_2))$   
FT: cuts-off the FT by a function of a spatial domain: low-pass

### Question 5: Time is of the essence (6 points)

A. (2) Why is stationarity a desirable property for a time series process?

With a stationary process, a longer time series (177) gives more info about the statistical properties (mean, autocorrelation etc) and they don't change

B. (2) Below is a plot of a monthly time series. The investigator who collected it expects there to be an annual seasonal pattern, but she is more interested in the long-term trend.



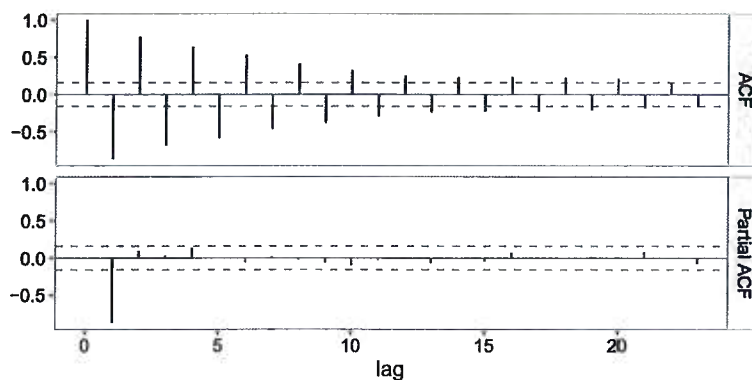
Suggest a method the investigator could use to model the seasonality, along with shortly motivating why.

More than one correct answer:

Examples:

- (a) calculate monthly means and model deviations from the mean
- (b) model the residuals of regression against the month

C. (2) Given the autocorrelation plot and partial autocorrelation plot below, what is your estimate for which kind of ARIMA model the researcher should attempt?



ARIMA(1,0,0) :

**Question 6: Small Dimensionality Reduction (5 points)**

We have a small database of ratings of 5 movies by 6 users (negatives are allowed, in the case that a user dislikes a movie much). They are represented by a  $6 \times 5$  matrix  $X$ , where each row corresponds to a user and each column corresponds to a movie.

$$\begin{array}{c} \text{u} \\ \text{s} \\ \text{e} \\ \text{r} \\ \text{s} \end{array} X = \begin{array}{c} \text{movies} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ -3 & -3 & -3 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & -1 & -1 \end{bmatrix} \end{array}$$

A.(1) What is the sample user mean, i.e. how is the average user rating movies?

obviously  $[0, 0, 0, 0, 0]$

B.(2) If SVD (Singular Value Decomposition) is applied, how many singular values you expect to have?

Explain why. What would change if two of the zeros were replaced by ones?

2 singular values (due to the structure of  $X$ )

If zeros are removed, that accounts for noise  
meaning that I expect more than  
2 (3 or 4)

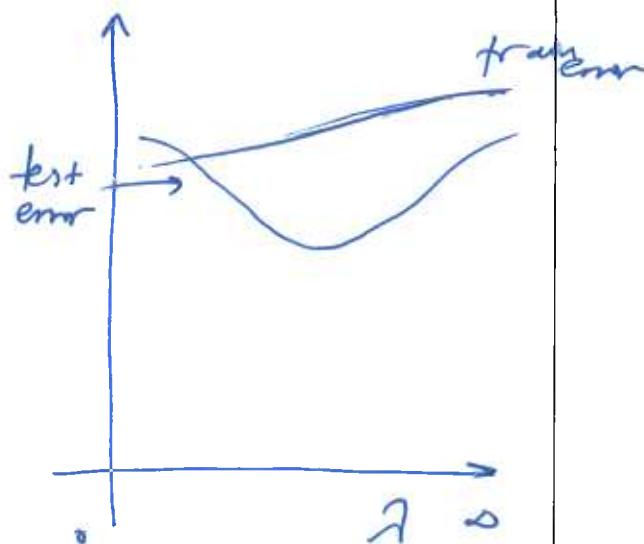
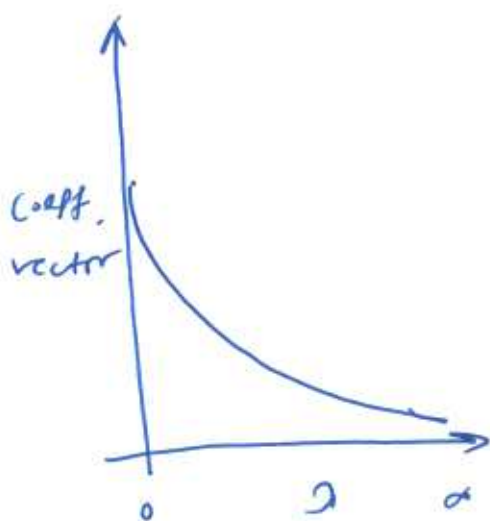
C.(2) A student very eager to apply a recommendation system on the above data, decided to do a per-user normalization (i.e. find the average rating of each user and subtract it from the values). Explain why the student decided to do this pre-processing step and what might be the problem in the dataset above.

- Remove rating bias for each user

- Problem because everything will become 0

### Question 7: Regularization magic (6 points)

A.(3) Suppose you are doing linear regression using an L2 regularization penalty. Sketch the average cross-entropy training and testing error you would expect to see when you run the algorithm on a dataset, as well as the norm of the coefficient vector, as a function of the regularization parameter  $\lambda$  (make one graph for the error curves and another for the coefficient vector norm). No explanation is necessary.





**B.(3)** Suppose you used cross-validation to find a good value for  $\lambda$ , on a dataset of 500 examples with 100 features. But, you have now acquired more data, so instead of the original 500 examples, now you have 50000. Would the same value of  $\lambda$  work best, or would you expect a higher or lower value of  $\lambda$  to work better? Justify your answer.

I expect a lower value of  $\lambda$

$\lambda$ : gives a bias-variance trade-off

data volume  $\uparrow$ , bias  $\downarrow$ , no need for high  $\lambda$

**Bonus (Meta)-Question (1 point)**

Find a mistake (typo, syntax or other) in this exam or the course slides.

---

Student name:

Student ID:

Page 15 of 15

Data Analysis 2017/2018

---

**Extra answer sheet.**