BA exam WS 20:21 - BA exam

Business Analytics (IN2028) (Technische Universität München)

Chair of Decision Sciences & Systems
Department of Informatics
Technical University of Munich

[redacted]

[redacted]

Compliance to the code of conduct
I hereby assure that I solve and submit this exam myself under my own name by only using the allowed tools listed below.

Signature or full name if no pen input available

Business Analytics: Endterm Exam

[redacted]

Working instructions

- This exam consists of **22 pages** with a total of **6 problems**.
- The total amount of achievable credits in this exam is 90 credits.
- Allowed resources: open book.
- Subproblems marked by * can be solved without results of previous subproblems.
- **Answers are only accepted if the solution approach is documented.** Give a reason for each answer unless explicitly stated otherwise in the respective subproblem.
- Do not write in red or green colors.
- Any intermediate or final numbers in your calculations may be **rounded to four (4) digits**.

[redacted]

ONLINE SUBMISSION

[redacted]

IN-b

[redacted]

Regression Analysis (15 credits)

You collected the following data about $n = 73$ different locations in Germany for a representative winter day in 2020:

Variable	Range	Explanation
snowfall	$\{0, 1\}$	occurrence of snowfall
temperature	$\mathbb{R}$	air temperature [°C]
humidity	$\mathbb{R}_{\geq 0}$	absolute humidity [g/m^3]
speed	$\mathbb{R}_{\geq 0}$	wind speed [km/h]
direction	$\{North, East, South, West\}$	wind direction

You would like to model *snowfall* by applying a logistic regression.

0 1 2

a)* Briefly describe the data set:

1. Name the dependent and independent variables.
2. What scale of measurement does the variable *direction* belong to? How would you process the variable *direction* for the regression?

0 1 2

b)* You decide to remove the variable *direction* since you consider it too unspecific. The remaining variables serve as an input to the logistic regression. Provide a formula to compute the likelihood function. The formula should include the variable names.

[redacted]

You apply a logistic regression and obtain the following output:

Residuals:

Min	1Q	Median	3Q	Max
-0.7088	-0.3151	-0.0666	0.3791	0.6938

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.808268	0.334746	-2.415	0.01841	*
temperature	-0.046228	0.010961	-4.218	7.37e-05	***
humidity	0.056305	0.019977	2.818	0.00629	**
speed	0.005037	0.005824	0.865	0.39010	

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null deviance: 15.753 on 72 degrees of freedom
Residual deviance: 10.861 on 69 degrees of freedom

c)* Interpret the estimated model:

0 1 2 3 4

1. Which variables are statistically significant at 5%? Explain your reasoning.
2. Interpret the *intercept* and the coefficient for *temperature*. Refer to the odds and provide necessary calculation steps.

[redacted]den R^2 for this model is 0.329482. Interpret this value. Explain the meaning of the McFadden R^2 and how it differs conceptually from the R^2 measure for OLS regression.

1
2
3

You are advised to include air pressure as another covariate. Introducing the variable *pressure*, your regression yields the following results:

```
Residuals:
Min        1Q         Median     3Q        Max
-0.74349   -0.28478   -0.05542   0.34002   0.67652

Coefficients:
            Estimate   Std. Error   t value   Pr(>|t|)
(Intercept) -0.798156  0.335005     -2.383    0.02000   *
temperature -0.128528  0.084762     -1.516    0.13407
humidity    0.056654   0.019986     2.835     0.00604   **
speed       0.004232   0.005884     0.719     0.47442
pressure    0.083481   0.085255     0.979     0.33096
—
Signif. codes: 0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1
Null deviance: 15.753 on 72 degrees of freedom
Residual deviance: 10.710 on 68 degrees of freedom
```

The new McFadden R^2 is 0.330173.

0
1
2
3
4

e)* Analyze these results:

1. Examine the significance and the McFadden R^2. What is the impact of adding the variable *pressure*?
2. Explain these results. Which Gauss-Markov property is affected and how can you test for it?

Problem 2 **Data Preparation (10 credits)**

You have obtained a dataset with metadata about 2,731 videos uploaded to a popular streaming platform. Some videos get featured on the platform's main page which usually helps video creators to reach a much wider audience. Based on your available data, you want to use a general linear model in R, i. e. a logistic regression, and perform statistical inference to determine the main drivers behind the platform's decision whether or not to feature a video.

While looking at the data, you realize that there are several data preparation steps you will need to undertake before you can run your model. Below, you are given three resources: (A) a short description of the variables in your data, (C) a summary of each column's distribution, as well as (B) the correlation between the numerical variables.

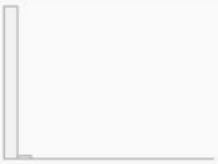
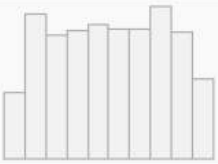
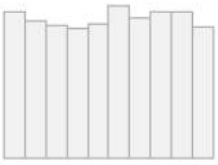
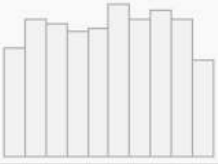
Resource A Your dataset has the following columns (see table below for types and distributions):

1. **video_id:** A unique identifier of each video.
2. **views:** The number of views the video has received on the platform.
3. **rating:** The star-rating (from 1 star '*' to 5 stars '*****').
4. **earnings:** earnings in US$ that were paid out to the video creator as part of a revenue-sharing agreement with the platform.
5. **account_gender**: the gender of the video creator
6. **account_age**: the age (in years) of the video creator
7. **sentiment_english**: A score between 0.0 (very negative comments) and 1.0 (very positive comments) that describes the sentiment of user comments posted below the video in English language.
8. **sentiment_other**: A score between 0.0 (very negative comments) and 1.0 (very positive comments) that describes the sentiment of user comments posted below the video in languages other than English.
9. **featured**: Whether or not the video was featured on the home page or not.

Resource B The *correlation matrix* between numerical features is given below:

	views	earnings	account_age	sentiment_english	sentiment_other
views	1.00000000	0.76582721	NA	0.01824006	0.01940797
earnings	0.76582721	1.00000000	NA	0.01947576	0.02048991
account_age	NA	NA	1	NA	NA
sentiment_english	0.01824006	0.01947576	NA	1.00000000	0.99860600
sentiment_other	0.01940797	0.02048991	NA	0.99860600	1.00000000

[redacted]: Summary of each variable's distribution

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	video_id [character]	1. 00Etx6TK 2. 016AmFYV 3. 01nrJJBC 4. 025OtRki 5. 02Kf4xNZ 6. 02VPhDIb 7. 03Xe0Uza 8. 06PdjMvm 9. 07yFfne3 10. 082cBmZJ [2721 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 2721 (99.6%)		2731 (100%)	0 (0%)
2	views [numeric]	Mean (sd) : 9300.9 (28084) min < med < max: 26 < 2824 < 799669 IQR (CV) : 7080 (3)	2351 distinct values		2731 (100%)	0 (0%)
3	rating [character]	1. * 2. ** 3. *** 4. **** 5. *****	543 (19.9%) 538 (19.7%) 533 (19.5%) 561 (20.5%) 556 (20.4%)		2731 (100%)	0 (0%)
4	earnings [numeric]	Mean (sd) : 194.3 (679.2) min < med < max: 0 < 39.7 < 19567 IQR (CV) : 147.5 (3.5)	2230 distinct values		2731 (100%)	0 (0%)
5	account_gender [character]	1. F 2. M	1236 (51.8%) 1152 (48.2%)		2388 (87.44%)	343 (12.56%)
6	account_age [integer]	Mean (sd) : 40.9 (13.2) min < med < max: 18 < 41 < 64 IQR (CV) : 23 (0.3)	47 distinct values		2388 (87.44%)	343 (12.56%)
7	sentiment_english [numeric]	Mean (sd) : 0.5 (0.3) min < med < max: 0 < 0.5 < 1 IQR (CV) : 0.5 (0.6)	2731 distinct values		2731 (100%)	0 (0%)
8	sentiment_other [numeric]	Mean (sd) : 0.5 (0.3) min < med < max: 0 < 0.5 < 1 IQR (CV) : 0.5 (0.5)	2731 distinct values		2731 (100%)	0 (0%)
9	featured [character]	1. No 2. Yes	2599 (95.2%) 132 (4.8%)		2731 (100%)	0 (0%)

[redacted]

[redacted]

Based on the resources and the goal above, identify **five problems** of the data set in its current form. For each, quickly describe why it is problematic in the given context, and name a possible data preparation strategy to overcome it. *After performing the steps you describe, the resulting data set should be suitable to fit your model given the following specification in R:*

```
model <- glm(featured ~ ., data=video_data_clean, family='binomial')
```

1
2
3
4
5
6
7
8
9
10

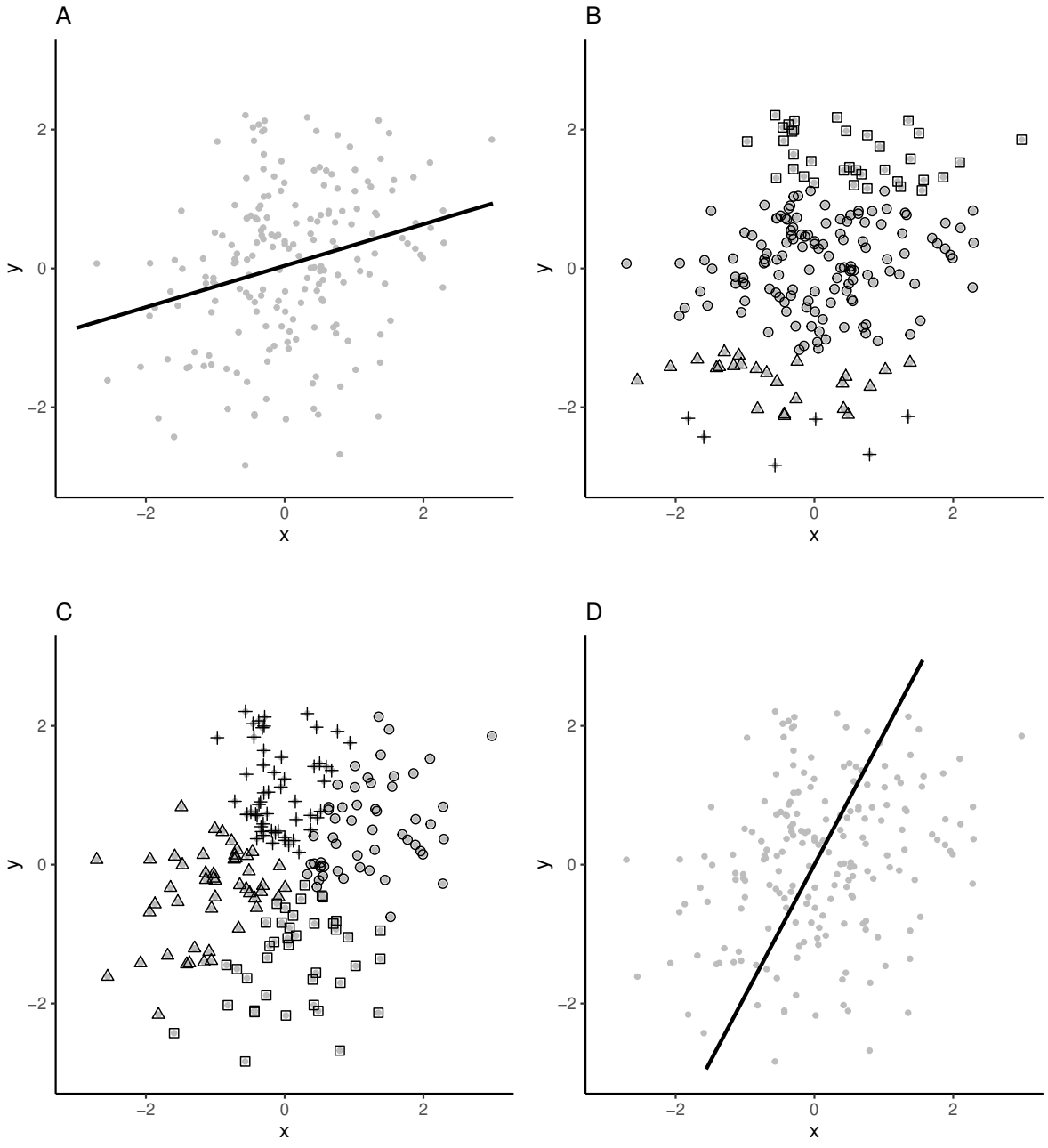
[redacted] 3 Identify Models (10 credits)

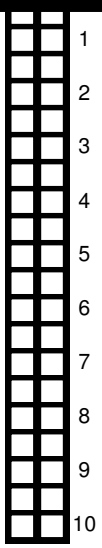
You have been hired by a client to help them understand a problem in one of their business processes. In your analysis, you have collected a dataset of 200 problem instances and fitted several statistical models on the data:

- **linear regression (OLS)**,
- **k-means**,
- **principal components analysis**,
- **decision tree**.

You are about to present your results to the clients and have printed handouts with visualizations of your models, but suddenly you realize that your assistant has forgotten to label each visualization with its title.

For each printout (A,B,C,D) below, (a) name which of the models is shown, (b) quickly summarize the visualization (What aspect(s) of the model are shown?) (c) justify your choice. (Why is it this model and not one of the others?)





Problem 4 Evaluation (18 credits)

Your cybersecurity company has developed two new email spam filters. The managers would like to have an analysis on both software systems (A and B). You have received a small but representative trial of experimental results:

True class	Probability of Classifier A	Probability of Classifier B
+	0.81	0.66
-	0.25	0.17
-	0.79	0.77
+	0.68	0.55
+	0.91	0.88
-	0.89	0.61
-	0.54	0.21
+	0.88	0.85
-	0.17	0.08
+	0.77	0.71

Each row represents an email, which is either spam (positive label) or no spam (negative label). The probabilities indicate the likelihood of being classified as spam by the respective software. The default cutoff value for both systems is 0.76.

0 1 2

a)* Calculate the *accuracy* of both software systems. Provide formula and calculations.

b)* Provide formulas and calculate *precision* and *specificity* for both software systems. Explain the results and characterize both software systems accordingly.

0 1 2 3 4

c)* Many customers prefer that a high share of the spam emails are classified as spam. Which software would you recommend? Name and calculate a meaningful metric.

0 1 2

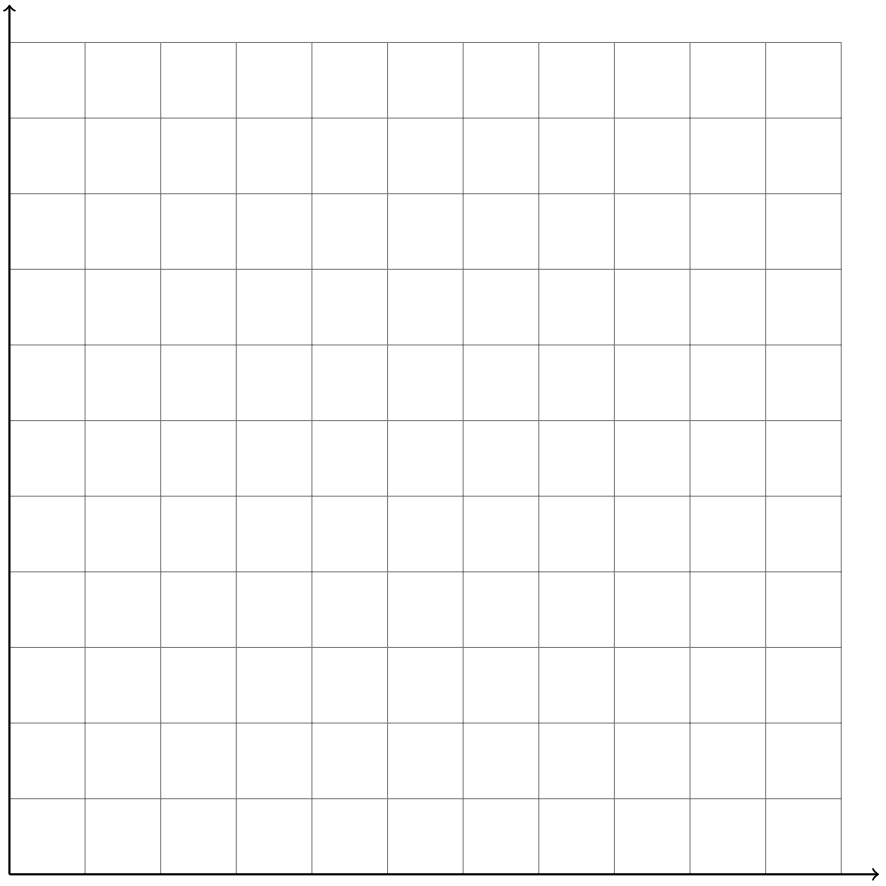
[redacted]

Note: Table identical as above (depicted here for ease of readability).

True class	Probability of Classifier A	Probability of Classifier B
+	0.81	0.66
-	0.25	0.17
-	0.79	0.77
+	0.68	0.55
+	0.91	0.88
-	0.89	0.61
-	0.54	0.21
+	0.88	0.85
-	0.17	0.08
+	0.77	0.71

0 1 2 3 4 5 6 7

d)* Draw the ROC curve of software A. Name the axes and write down the (x,y)-coordinates for every point. Mark the point corresponding to the default cutoff value of 0.76. Would you argue that this cutoff value is appropriate? Explain your reasoning.



[redacted]

e)* Do you agree with the following statement? Explain your reasons.

"The specificity can be easily computed from the ROC curve. For a particular cutoff value, it equals the slope of the line connecting the origin and the point on the ROC curve corresponding to the cutoff value."

0 1 2

Problem 5 Decision Trees (17 credits)

A laundry manager wants to optimize her energy and detergent usage. Therefore, she tested the washing results for various settings and reported her results in Table 5.1.

Temperature	Detergent	Dirt	Result
30	low	low	good
30	low	high	bad
30	low	medium	bad
30	high	high	bad
40	low	low	good
40	low	medium	bad
40	high	high	bad
40	low	medium	good
60	high	medium	good
60	high	medium	good

Table 5.1: Laundry results.

0 1 2 3 4 5

a)* Of the first three attributes, find that one that would be at the root of a decision tree using the **information gain** criterion. For the numeric attribute *Temperature* consider binary splits.

Note: For this exercise, we will again use $0 \cdot \log(0) = 0$.

b)* Assume there is an additional data point with a missing value:

Temperature	Detergent	Dirt	Result
40	low	?	good

However, your colleague had access to all data (including the new observation) and calculated a **gain ratio** of 0.1569 for the attribute *Dirt*. What is the missing value marked by the question-mark?

0 1 2 3 4 5 6

c)* What is the reason for pruning and how does the pruning method subtree replacement work?

0 1 2

d)* Generally, why is it not a good idea to discretize numerical attributes into many bins?

0 1 2

[redacted]

0
1
2

e)* You build a decision tree with pruning activated. Nevertheless, it is still large and difficult to interpret. What can you do to make it smaller?

[redacted]

[redacted]

Problem 6 Inference and Causality (20 credits)

The following scenario is based on real events, however numbers and other details are adjusted for academic purposes.

In 1933, the prohibition in the US ended and alcohol could be consumed legally. Most of the states in America (among them South Dakota) chose to set the minimal legal drinking age to 21 but some (among them North Dakota) set it to 18, arguing that people are adults with 18 and it could diminish the thrill for it. With the years the federal government aimed to have a legal drinking age of 21 in all states and therefore, starting in 1984 withheld federal aid for states in which the minimum legal drinking age was still 18. By 1988 all 50 states had a minimum legal drinking age of 21. North Dakota changed the drinking age in 1984. You have a dataset of the years 1980 to 1988 with the yearly death rates caused by car accidents in North and South Dakota.

In a study you want to address the following question: Does increasing the legal drinking age from 18 to 21 decrease the number of deaths caused by car accidents?

a)* What type of experiment/study does the above described scenario allow for? Provide 2-3 arguments.

0 1 2 3

b)*

1. What technique can you apply to identify causal effects in this setting? Explain the technique and argue briefly why you choose it (3-4 sentences). (6P)

2. Why do you need that technique and cannot simply make a causality statement based on the data of North Dakota only? (3P)

0 1 2 3 4 5 6 7 8 9

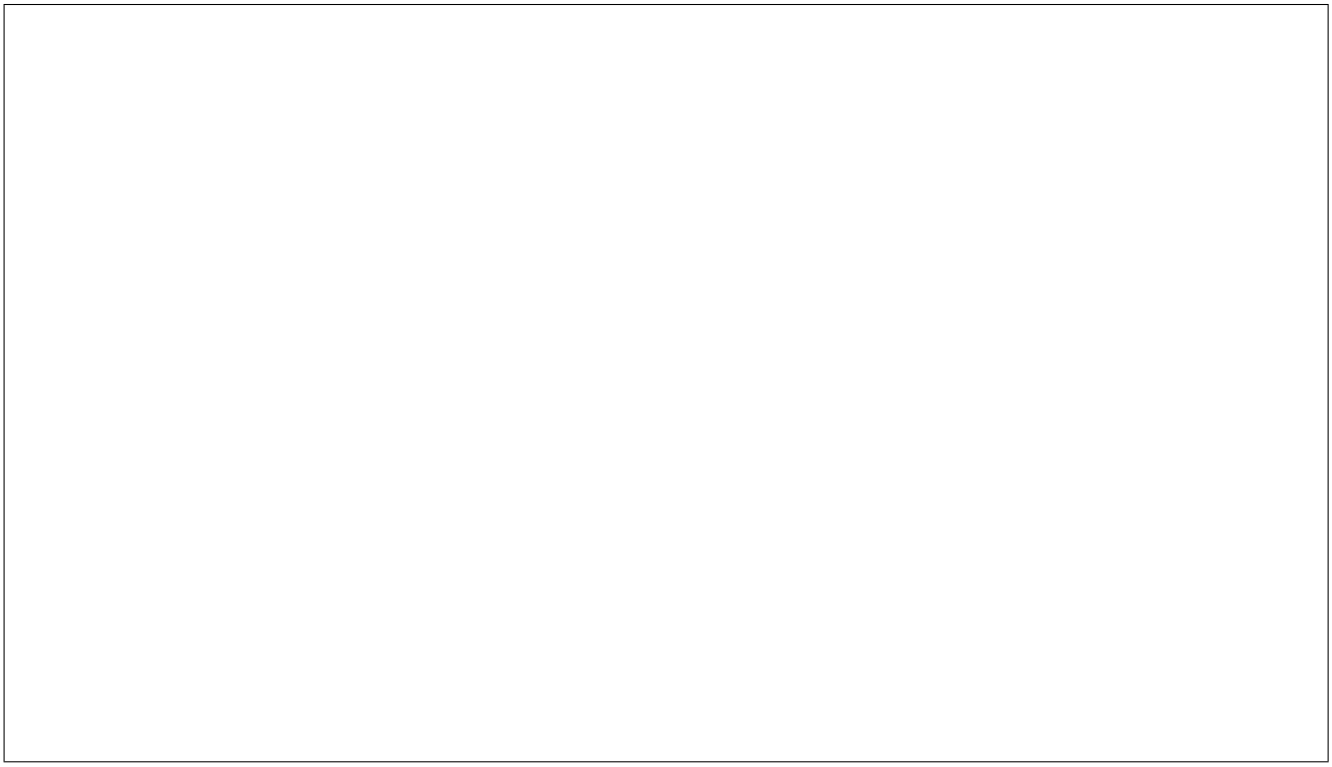
c)*

Fictional

Assume you have data of 3,000 drivers in North Dakota in 1970. Some of them voluntarily performed a driver safety training in that year. Additionally you know for each of the drivers whether they have been involved in a car accident within in the next 5 years.

Now, you want to make a statement of whether the driver safety training helped the drivers to avoid car accidents.

1. Name one appropriate technique to help you identify causal effects in this setting. (2P)
2. Explain in 2-3 sentences why it is appropriate. (3P)
3. Describe in 2-3 sentences how you would apply it. (3P)



Additional space for solutions–clearly mark the (sub)problem your answers are related to and strike out invalid solutions.