Guidelines:

Research design & term paper draft

Paul C. Bauer

mail@paulcbauer.de

First Version: February 2019

This Version: February 9th 2022

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Please write your document with Quarto: Check out [this template](https://osf.io/ur4xn)!

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The term paper draft should be around two pages long and answer a set of questions. Below are the relevant questions for term papers that either focus on a causal or a predictive question.

**Table of contents**

[**1 Draft for a Causal Research Question**](#_cvqey0ywcmnf) **2**

[**2 Draft for a Predictive Research Question**](#_cltpdl2l5rfw) **4**

# 

# 1 Draft for a Causal Research Question

The questions below should structure the draft for a term paper that answers a causal research question.

* **Research question:** What is the question you would like to answer?
  + e.g., Is there a causal effect of education on income?
  + e.g., Is there a causal effect of slave trade on trust?
  + e.g., Is there a causal effect of minimum wage on employment?

*Precise research questions should already specify the population one is interested in, e.g., “Is there a causal effect of education on income among German citizens (or German residents)?”*

* **Ideal experiment:** What would be the ideal experiment to study this question?
* **Population (Target):** What is the population you are making an inference about?
  + e.g., humans, German citizens, refugees, all countries, cities
* **Sample:** What kind of sample are you using?
  + e.g., [simple random sample](https://en.wikipedia.org/wiki/Simple_random_sample) of Germans, a random sample of cities
  + Subquestions: What is your [sampling frame](https://en.wikipedia.org/wiki/Sampling_frame)? How was your sample collected?
    - e.g., called individuals randomly drawn from a phone book, contacted individuals randomly drawn from an address list provided by the government, opt-in online sample invited through email
* **Unit:** What is the fundamental unit you are analyzing/comparing?
  + e.g., individuals, cities, etc. (see research question)
  + Important: Make a clear distinction between individual units (e.g., individuals) and aggregates of units (e.g., individuals in certain cities) as well as observations (e.g., we may observe the same individual across time). Think about on which level the treatment is operating.
* **Data/Observations:** What data are you using (*e.g., cross-sectional, panel, etc.*)? When was the data collected?
  + ..if different variables have been collected at different time points indicated that here (*e.g. victimization measures in 2005, trust measured in 2010*)
* **Measurement:** How do you measure outcome, treatment and control variables?
  + e.g., income measured through a survey question
  + Subquestions: What do these measures look like? Survey questions? Scales/Categories? Coding decisions etc.? When are they measured? Do you have timestamps for changes?
  + How do you construct the respective variables?
    - e.g., *For the treatment variable I take the education scale (0-5) and construct a new variable (0,1) where 0 is educational levels lower than and including high school, 1 is educational levels higher than high school.*
* **Measurement error:** Do you expect any [measurement error](https://en.wikipedia.org/wiki/Observational_error)?
  + e.g., the outcome is income, and males will report higher than actual income
  + Subquestions: If yes, for all units or particular units? What is the direction of that error - under- or overestimation?
* **Identification strategy/design:** What identification strategy are you using?
  + e.g. randomized experiment, field experiment, observational data using certain method/assumptions
    - e.g., I use cross-sectional data. I control for X1, X2, X3 and assume that assignment to values of D is random conditional on these covariates (“selection on observables”)
* **Model & Unit of analysis & Estimation:**
  + How do you estimate the causal effect(s)? What model are you using?
    - e.g. OLS regression, Comparison of means with a t-test
  + What is the unit of analysis in your model?
    - e.g. individuals, individuals\*time, countries\*years
  + Which causal quantity are you interested in?
    - e.g. ATE, ATT etc.
* **Theory/Causal Mechanism:** What is the “assumed” causal chain/mechanism that connects values of D with values of Y?
  + e.g., A reaches a higher educational level, A looks for better-paid jobs, A gets a job that is better paid...
* **Hypotheses:** What effect (direction, size) do you expect for D on Y? Why? What theories is your expectation based on?
  + “The higher/lower X, the higher/lower Y”
  + Think of the hypothesis in terms of counterfactuals; If Peter had a higher education level (3 instead of 2), he would have a higher level of income (1000 instead of 500); If the the individuals in the control group (0) had been in the treatment group (1), they would have…
* **Previous evidence:** Is there previous evidence on the question you study?
  + If there is any previous research, provide a list of studies/books that may be relevant to your research question.
  + Search <https://scholar.google.ch/> for the concepts you study (e.g., happiness etc.). Make sure that you also search for concepts closely linked to the ones you are interested in. For instance, research on happiness can be also found under the label well-being etc.

# 2 Draft for a Predictive Research Question

The questions below should structure the draft for a term paper that answers a predictive research question.

* **Research question:** What is the question you would like to answer?
  + e.g., Can we predict Y (e.g., income)?
  + e.g., How well can we predict Y?
  + e.g., Which ML model should be used to predict Y?
  + e.g., What features do we need to predict Y?

*Precise research questions should already specify the population one is interested in, e.g., “How well can we predict income/victimization among German citizens (or German residents)?”*

* **Population (Target):** What is the population in which you predict the outcome?
  + e.g., humans, German citizens, refugees, all countries, cities
* **Sample:** What kind of sample are you using?
  + e.g., [simple random sample](https://en.wikipedia.org/wiki/Simple_random_sample) of Germans, a random sample of cities
  + Subquestions: What is your [sampling frame](https://en.wikipedia.org/wiki/Sampling_frame)? How was your sample collected?
    - e.g., called individuals randomly drawn from a phone book, contacted individuals randomly drawn from an address list provided by the government, opt-in online sample invited through email
* **Unit:** What is the fundamental unit for whom you are predicting Y?
  + e.g., individuals, cities, etc. (see research question)
  + Important: Make a clear distinction between individual units (e.g., individuals) and aggregates of units (e.g., individuals in certain cities) as well as observations (e.g., we may observe the same individual across time). Think about on which level your outcome variable is operating.
* **Data/Observations:** What data are you using (*e.g., cross-sectional, panel, etc.*)? When was the data collected?
  + ..if different variables have been collected at different time points indicated that here (*e.g., income measures in 2010, education measures in 2005*)
* **Measurement:** How do you measure the different variables (your outcome and your predictors)?
  + e.g., income measured through a survey question
  + Subquestions: What do these measures look like? Survey questions? Scales/Categories? Coding decisions etc.? When are they measured? Do you have timestamps for changes?
  + How do you construct the respective variables?
    - e.g., *For the treatment variable I take the education scale (0-5) and construct a new variable (0,1) where 0 is educational levels lower than and including high school, 1 is educational levels higher than high school.*
* **Measurement error:** Do you expect any [measurement error](https://en.wikipedia.org/wiki/Observational_error)?
  + e.g., the outcome is income, and males will report higher than true income
  + Subquestions: If yes, for all units or particular units? What is the direction of that error - under- or overestimation?
* **Machine learning method:** What kind of general machine learning method are you using for your predictions?
  + e.g., logistic regression, linear regression, random forests, deep learning, etc.
* **ML Model & Unit of analysis & Estimation:**
  + How do you produce the predictions? What model are you using exactly?
    - e.g., random forest with x features etc.
  + What is the unit of analysis in your model?
    - e.g., individuals, individuals\*time, countries\*years
* **Theory/Mechanisms:** Is there a mechanism that connects your features/predictors with your outcome? Or are What is the “assumed” causal chain/mechanism that connects values of D with values of Y?
  + e.g., A reaches a higher educational level, A looks for better-paid jobs, A gets a job that is better paid...
* **Training/testing:** How do you proceed in training/testing the ML Model?
* **Assessment of accuracy:** How do you assess the quality of your predictions?
* **Previous studies:** Are there previous papers that built a machine-learning model for the prediction problem? What are they? How do they proceed?

# 