Problem set 1

Yonatan Schoen

2023-04-06

ML for economics - Problem Set 1

Part 1 - open questions

Question 1

As we discussed in the first lecture, ML models typically prioritize predictive accuracy over explanation, and specifically casual one. While this may be sufficient for some applications, it can be problematic when trying to understand the underlying mechanisms of economic systems and design policies that can improve them. Economics is not just about prediction but also about understanding the causal relationships between different variables.

This issue can be even more problematic concerning the fact that economics models usually involve complex interactions between many variables, limited and noisy data, and policy trade-offs that are challenging to capture in ML models. Its also worth mentioning that economic issues often involve ethical considerations and policy trade-offs, which are challenging to capture in ML models (just think about designing policies that balance economic growth and environmental considerations, those requires more than just predictive skills but also normative point of view).

Question 2

Linear regression models can indicate correlations between different variables, but as we know well from economics studies, correlation does not necessarily imply causation. To prove causality, we need to carefully design experiments that meet certain criteria based on initial assumptions.Specifically, it must be the case where the cause (in most cases the “treatment” variable) happens before the effect and secondly that the relationship between the cause and effect isn’t explained by some other factor.

In addition, establishing causality requires us to find good “identification” of the casual effect in real life. We can of course define careful experimental design, including random assignment and control groups, to ensure that any observed effects are due to the treatment and not other factors. Nevertheless, it is not easy to achieve this level of experimental control in many economic contexts, which makes it challenging to establish this kind of interfrence.

Unfortunately, it’s not always easy to design experiments like this in economic contexts and moreover in ML designs. As mentioned above, many machine learning models prioritize making accurate predictions over being easy to interpret. This can make it harder to use these models to establish causality because they may show us relationships between different factors, but we can’t always be sure if those relationships are causal. So while machine learning can be useful in economics, it’s important to use other tools and methods to help us understand causality.

Question 3

a.) The linearity assumption means that the relationship between the dependent variable (Y) and independent variables (X's) is linear - i.e. straight and constant. This allows us to estimate the effect of each X on the dependent variable (as stright line with constant slope), and to make easy predictions, based on the linear form.

b.) The separability assumption means that the effect of each independent variable (X's) on the dependent variable (Y) is not affected by the values of the other independent variables. This is pretty strong assumption, which means that our independent variables do not interact with each other in their effect on the Y. Nevertheless, this assumption simplifies the model and makes it easier to interpret, but as mentioned it may not always be realistic in practice.

c.) The notation X~N(μ,σ2) is used to describe a normal distribution of a random variable X, where μ and σ^2 represent the mean and variance of its distribution, respectively. In our economical context, normal distributions are commonly used because many real-world phenomena have properties that can be approximated by normal distributions (i.e. central limit theory). This means that we can use statistical techniques to understand and analyze the data, and to make predictions about the behavior of the phenomenon we are studying.

Part 2 - R and Rstudio

I already have [R] and Rstudio on my computer.

Question 1 - load the required packages

```
library(tidyverse)
library(kableExtra)
library(tinytex)
```

Question 2 - manipulate the iris data frame

```
data(iris)
iris %>%
  select(c(Sepal.Length, Sepal.Width, Species)) %>%
  group_by(Species) %>%
  summarise(Average_Sepal_Length = mean(Sepal.Length)) %>%
  kbl() %>%
  kable_styling()
```

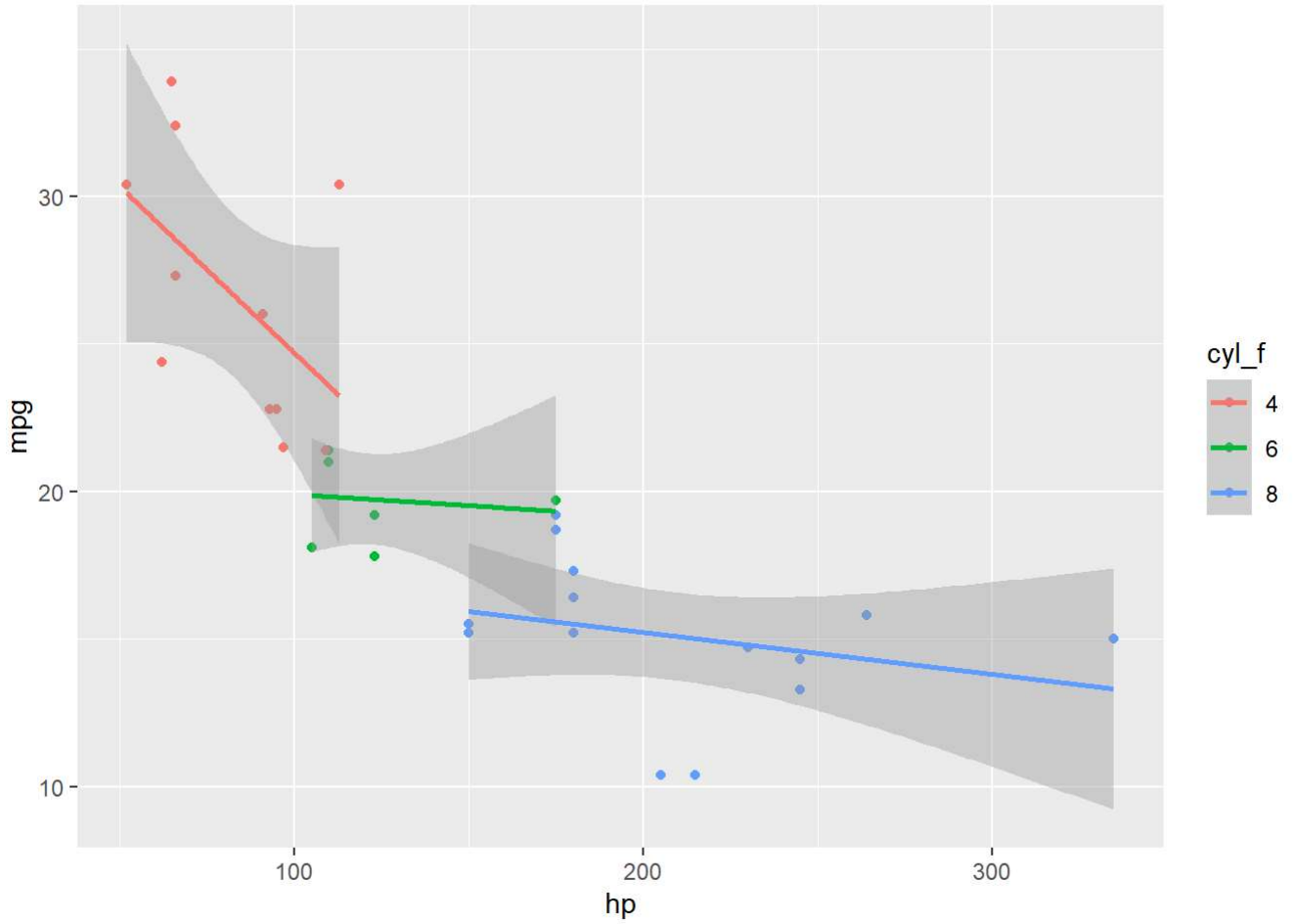
Species	Average_Sepal_Length
setosa	5.006
versicolor	5.936
virginica	6.588

Question 3 - plot mtcars using ggplot

```
data(mtcars)

mtcars %>%
  mutate(cyl_f = as.factor(cyl)) %>%
  ggplot(aes(x = hp, y = mpg, color = cyl_f)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Part 3 - Git and Github

Question 1 - Git configuration

(I didnt understand this part exactly, used the code from the solutions that were given)

```
library(usethis)

use_git_config(
  scope = "project",
  user.name = "Yonatan",
  user.email = "yonatanschoen@gmail.com"
)
```

Questions 2 and 3 done as well.

I'll be happy for some explanation about how "usethis" works. Because I already synced my R project to github so thats all of my projects updated automatically when saving.

Thanks.