ProbSet 1, January 16

Gideon Gordon gideongordon2029@u.northwestern.edu

Due: January 16, 2026

Submission: https://canvas.northwestern.edu/courses/245562/assignments/1676620

0.1 Problem 1

Explain in your own words: What is regression trying to achieve? How does it differ from simply calculating the correlation between two variables?

Linear regression is basically trying to identify the "best linear predictor", so, the line which comes closest to all the points, controlling for some other variables. This is a roundabout way of approximating the unobservable Conditional Expectation Function of y conditional on every possible value of some x.

The BLP is something we want mostly because it's straightforwardly interpretable, not because the CEF is actually a straight line. We can also apply some transformations to the BLP to model more complicated hypothesized CEFs.

This is all conceptually distinct from a simple correlation between x and y in that calculating a regression doesn't try to find the CEF, and practically distinct in that a regression can take forms aside from a single coefficient (because it can be nonlinear on occasion), has an intercept (because it's a line), and can control for other variables.

0.2 Problem 2

In your experience and judgment, what is the difference between a good and a bad scatterplot? What are the goals of visualizations like scatterplots in the context of social science data analysis?

The goal of a visualization is to give some vividness to analysis and inspire thinking about possible patterns. To do either of those things, it needs to be tailored to human beings, even if the math is fairly clear. A good scatterplot should be readable and human-accessible. If it's not, don't show the scatterplot, just give me a line or a table.

This means, first of all, individual dots should be easily interpretable as individual cases of the phenomenon of interest. The exact opposite is what we saw on the democracy-GDP country-year scatterplot. Each dot there was a country-year, which is a weird way to conceptualize a phenomenon we believe occurs at the country level.

Second, any data shown by individual dots should be readable with the unaided human eye. A possible failing here that I see in some scatterplots is when authors decide to include labels for all the points, or use labels as points on the scatterplot. Often, when data points are close together, the labels overlap and only a few specific points are readable. This is totally useless to the reader.

Third, the central tendency of the data should be visible to an average human. This has two parts. First of all, showing a big blob of points is useless, even if based on your regression there is a significant central tendency worth paying attention to. If a human being cannot look at the scatterplot and say something like "huh, all the countries with civil wars cluster over here, and all the countries without civil wars cluster over here", the scatterplot is ineffective.

A second element of making the central tendency visible is to be aware of human biases towards outliers in the data. As we discussed with the democracy-GDP data, our eyes are

drawn automatically to a few country-years with very high GDP and very autocratic regimes. This can be corrected in part by including a regression line on the scatterplot or by including some note explaining the outliers.

0.3 Problem 3

Let's figure out whether there have been equal numbers of ICE arrests in 2024 in states with Democratic and Republican governors. We have a dataset pulled from Wikipedia that lists details about governors, and the deportation data. Let's start by getting both of these loaded into our R workspace.

```
governor_data <- read.csv("https://github.com/jnseawright/ps210/raw/refs/heads/main/Data
ice_arrests <- read.csv("https://github.com/jnseawright/ps210/raw/refs/heads/main/Data/i
# Now what we want to do is create a new variable in the ICE data that records the parti
# governor where the arrest happens.

#This command creates a new empty variable called Party.
ice_arrests$Party <- NA

#This block of commands is going to loop through the ice_arrests database and
#check the relevant partisanship of the governor for each arrest
for (i in 1:nrow(ice_arrests)){
  #This command is checking if a given arrest happened in one of the 50 states.
  #Some arrests have no recorded location, some happen in international travel,
  #some happen on military bases, etc. For those, we'll record the party as
  #missing.
  if (!ice_arrests$State[i] %in% levels(as.factor(governor_data$State)))
    ice_arrests$Party[i] <- NA
    #When the party isn't missing, we'll set it from the governor data.
```

```
    else ice_arrests$Party[i] <- governor_data$Party[governor_data$State==ice_arrests$St
}
```

0.3.1 3a.

Run a regression predicting state-level ICE arrests relative to the party controlling the state governorship. Interpret the results.

So, what we'll want to do here is construct a chart of number of arrests by state. Then the question is: "Do states with Republican governors see more arrests than states with Democratic governors?"

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(dbplyr)
```

```
Attaching package: 'dbplyr'

The following objects are masked from 'package:dplyr':

    ident, sql
```

```r
library(ggplot2)
state_ice_arrests = ice_arrests |>
  group_by(State, Party) |>
  summarize(arrests = n())
```

```
`summarise()` has grouped output by 'State'. You can override using the
`.groups` argument.
```

This is more or less all the information we need. We don't care about anything else (so far).

```r
partisan_arrests = lm(data = state_ice_arrests, formula = arrests ~ Party)
summary(partisan_arrests)
```

```
Call:
lm(formula = arrests ~ Party, data = state_ice_arrests)

Residuals:
   Min     1Q Median     3Q    Max
 -5427  -3352  -1888    174  53937

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)         3738       1877   1.991   0.0522 .
PartyRepublican     1733       2555   0.678   0.5008
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9004 on 48 degrees of freedom
```

```
  (12 observations deleted due to missingness)
Multiple R-squared:  0.009497,  Adjusted R-squared:  -0.01114
F-statistic: 0.4602 on 1 and 48 DF,  p-value: 0.5008
```

That p-value is massive, and so I will say that nothing can be learned from this regression.

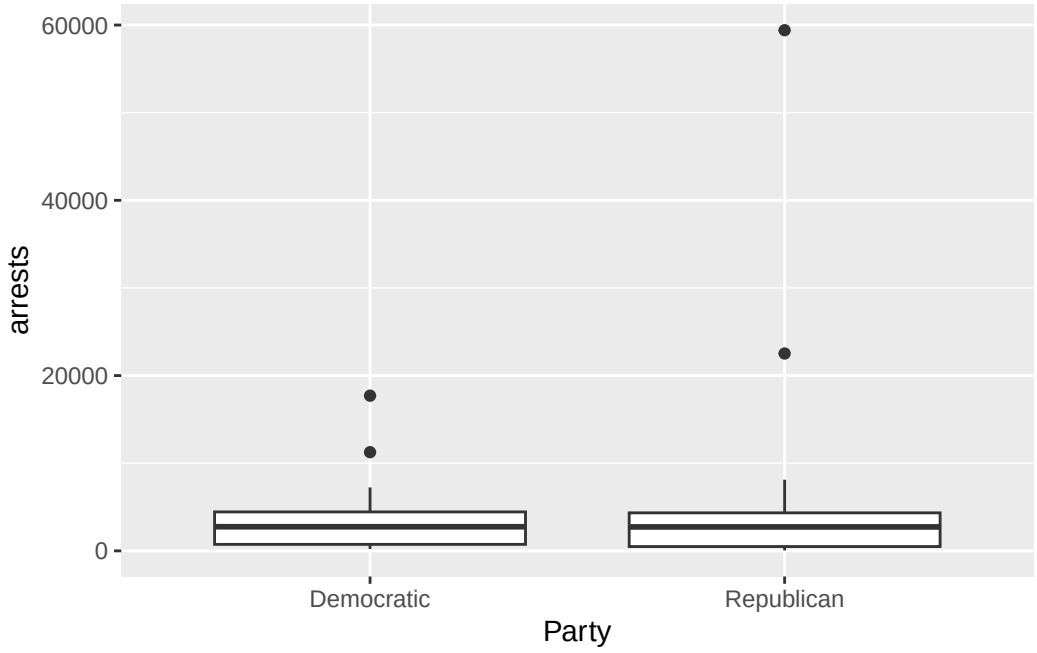
0.3.2 3b.

Before running your regression, create an appropriate visualization to explore the relationship between governor's party and ICE arrests. Consider using a boxplot, violin plot, or jittered scatterplot. Describe what you observe in the visualization and how it complements or contrasts with your regression results.

Hint: You may need to aggregate the data to state level or sample appropriately for visualization given the dataset's size.

I excluded every place for which no governor was available, then set up a boxplot.

```
ggplot(data = na.omit(state_ice_arrests), aes(x = Party, y = arrests)) +
  geom_boxplot()
```



This is a distinctly uninformative visualization, which perfectly complements my distinctly uninformative regression. We see that Republican and Democratic states are very similar in the absolute numbers of arrestees. There are some major outliers, which I can't safely ignore given how few states there are.

Also though, recall that Republican states tend to have smaller populations, so absolute numbers are not necessarily the best metric here.

0.4 Problem 4

A common feature of regression in social science applications is multivariate analysis. It might make sense to add state populations as a conditioning variable in the regression from the previous problem. Let's start by adding data on population to our ice_arrests dataset.

```
state_pops <- read.csv("https://github.com/jnseawright/ps210/raw/refs/heads/main/Data/st
#Now we can use a version of our code from above that copies in state populations
#instead of partisanship.

#This command creates a new empty variable called Population.
ice_arrests$Population <- NA

#This block of commands is going to loop through the ice_arrests database and
#check the relevant population of the state for each arrest
for (i in 1:nrow(ice_arrests)){
  #This command is checking if a given arrest happened in one of the 50 states.
  #Some arrests have no recorded location, some happen in international travel,
  #some happen on military bases, etc. For those, we'll record the population as
  #missing.
  if (!ice_arrests$State[i] %in% levels(as.factor(state_pops$State)))
```

```
    ice_arrests$Population[i] <- NA
    #When the population isn't missing, we'll set it from the state population data.
    else ice_arrests$Population[i] <-
        state_pops$Population2024[state_pops$State==ice_arrests$State[i]]
}
```

```
library(readr)
#This variable often reads in with commas and gets treated as text, so we'll
#make sure to convert it to an actual number.
ice_arrests$Population <- parse_number(ice_arrests$Population)
```

0.4.1 4a.

Run a regression predicting state-level ICE arrests as a function of which party controls the governorship and also state population. Once again, interpret your results, also describing any interesting comparisons with the bivariate results in Problem 3.

```
state_party_population = ice_arrests |>
  select(State, Population, Party)


state_ice_arrests_population = state_party_population |>
  group_by(State, Party, Population) |>
  summarize(arrests = n())
```

```
`summarise()` has grouped output by 'State', 'Party'. You can override using
the `.groups` argument.
```

```
partisan_population_arrests = lm(arrests ~ Party
                                 + Population,
                         data = state_ice_arrests_population)
```

```r
summary(partisan_population_arrests)
```

```
Call:
lm(formula = arrests ~ Party + Population, data = state_ice_arrests_population)

Residuals:
     Min       1Q   Median       3Q      Max
-15669.7  -1464.5   -295.6   1579.1  29785.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.861e+03  1.413e+03  -2.732  0.00883 **
PartyRepublican  3.934e+03  1.571e+03   2.504  0.01580 *
Population       9.444e-04  1.037e-04   9.111 5.92e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5471 on 47 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.6419,    Adjusted R-squared:  0.6267
F-statistic: 42.13 on 2 and 47 DF,  p-value: 3.303e-11
```

State population is the best predictor of absolute number of arrests, but a Republican party governor also matters a bit even controlling for population. The substantive effect of having a Republican governor is also fairly large, a difference of about 3,000 arrests.

This feels a little weird, though. Saying that having a Republican governor is correlated with an increase in arrests of 3,000 people, controlling for population, seems to suggest that

there's a flat increase, rather than a change in policies to allow more cooperation (simply by putting these variables in the equation, we're assuming there's some sort of quasi-causal story, even if we can't yet prove causation). I would expect a change in policies to create an increase in the *relative* number of arrests.

I'm going to restructure to look at arrests proportional to population (ICE arrests per ten thousand people)

```
state_ice_arrests_population = mutate(state_ice_arrests_population,
              proportional_arrests = (10000*(arrests/Population)))
```

Now instead of controlling for population, I'll run the regression straight. I don't think there will be too much of a difference but I'll feel better about it.

```
partisan_proportional_arrests = lm(proportional_arrests ~ Party,
              data = state_ice_arrests_population)
summary(partisan_proportional_arrests)
```

```
Call:
lm(formula = proportional_arrests ~ Party, data = state_ice_arrests_population)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9048 -2.5585  0.0258  1.9462 12.5446

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.5845     0.7306   6.275 9.52e-08 ***
PartyRepublican   1.8567     0.9942   1.867    0.068 .
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.504 on 48 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.06773,   Adjusted R-squared:  0.04831
F-statistic: 3.487 on 1 and 48 DF,  p-value: 0.06796
```

The significance of having a Republican governor falls a lot in this measurement, maybe because of how I rescaled things; the p-value is all the way to 0.1 now. Still, I prefer interpretation this way, since it gives me a number I can interpret better. Having a Republican governor rather than a Democratic governor is correlated ($p < 0.1$) with an increase in ICE arrests of 1.8 arrests per ten thousand people.

0.4.2 4b.

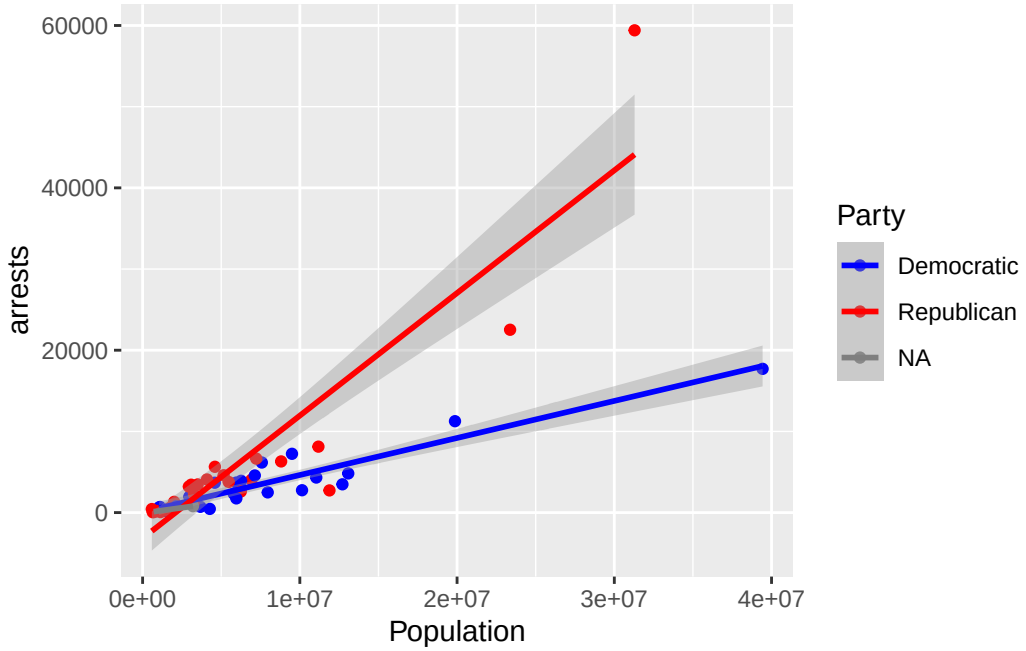
Create a scatterplot of ICE arrests against state population, using color to distinguish between states with Democratic and Republican governors. Add regression lines (either separate lines for each party or one overall line) to visualize the relationship. Discuss how this visualization helps you understand the multivariate regression results. What patterns do you see that might not be apparent from the regression table alone?

```
dems_and_repubs_charted_lm = ggplot(state_ice_arrests_population,
  aes(x = Population, y = arrests, color = Party)) +
  geom_point() +
  geom_smooth(method="lm") +
  scale_color_manual(values=c('blue', 'red', 'grey'))
```

(I had to modify the colors since otherwise the Republicans were in blue and the Democrats in red, contrary to America's most fundamental values).

```
dems_and_repubs_charted_lm
```

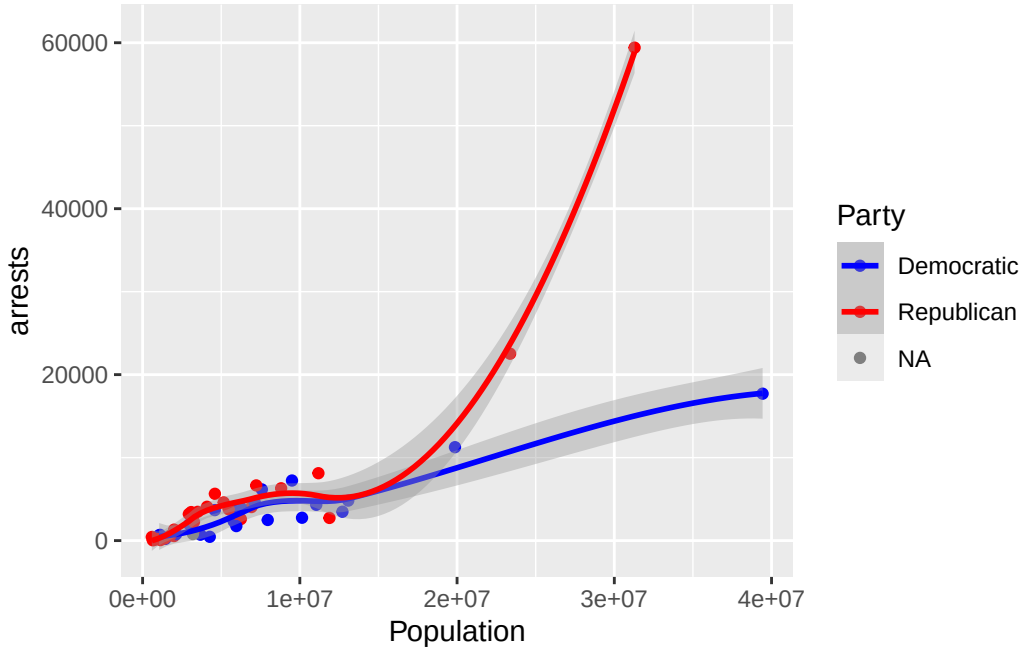
```
`geom_smooth()` using formula = 'y ~ x'
```



The regression table alone did not help me much with understanding how the partisan difference plays in practice. My initial evaluation showed that a governor being Republican increased arrests by about 3,000. But the effect is really most significant for the big states.

```
dems_and_repubs_charted_loess = ggplot(state_ice_arrests_population, aes(x = Population,
  geom_point() +
  geom_smooth() +
  scale_color_manual(values=c('blue', 'red', 'grey'))
dems_and_repubs_charted_loess
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

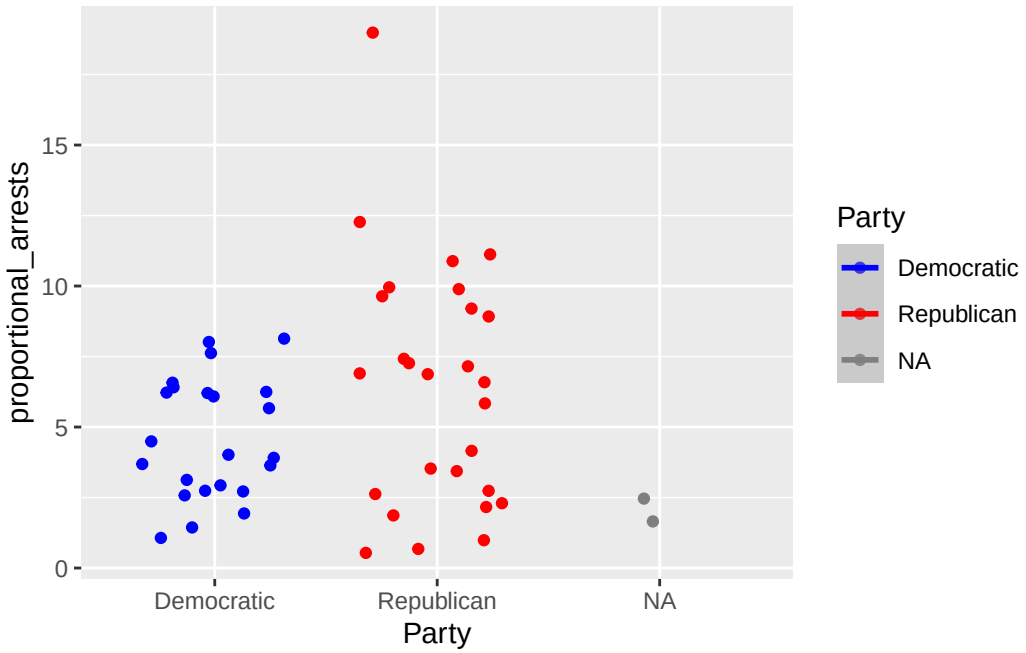


The effect is driven primarily by a few flagship states — I suspect Texas and Florida, the largest Republican states, are the most important ones. Most of the time, the difference between Republican and Democratic governors doesn't matter much. The role of these particular outliers is invisible in the overall regression.

To visualize just differences between Republicans and Democrats with arrests as a proportion of population"

```
dems_and_repubs_charted_proportionally = ggplot(state_ice_arrests_population,
  aes(x = Party, y = proportional_arrests, color = Party)) +
  geom_jitter() +
  geom_smooth(method="lm") +
  scale_color_manual(values=c('blue', 'red', 'grey'))
dems_and_repubs_charted_proportionally
```

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



0.5 Problem 5

Reflection question: Based on your analysis in Problems 3 and 4, what are at least two limitations of using governor's partisanship as a predictor of ICE arrests? Consider both methodological issues (e.g., measurement, confounding variables) and substantive concerns (e.g., causal mechanisms, political context).

On a methodological level, the first limitation is patterned differences between Republican and Democratic states overall. Democratic states may have strong networks of other organizations able to intervene to interfere with arrests, as we've seen in Minnesota. Electing a Democratic governor and preventing ICE arrests might both be caused separately by strong grassroots liberal activism including at the local government level. Republican states tend to be more rural and agricultural, which correlates with sectors where there is strong demand for undocumented labor — there may simply be more undocumented immigrants in farm states. If farmers tend to vote for Republicans at higher rates *and* are more likely to employ undocumented labor, governor partisanship and arrest outcomes could be spuriously correlated.

Substantively, one hypothesized causal mechanism would be that a Democratic governor is less likely to cooperate with ICE agents seeking arrestees than a Republican governor. We have seen that ICE targeting is highly heterogeneous within a single state, and especially since early 2025 has targeted specific selected cities — Los Angeles, Chicago, Charlotte, New Orleans, and now Minneapolis. ICE targeting here, even in Republican states like North Carolina and Louisiana, almost exclusively targets liberal cities with Democratic mayors.

We still might guess that Republican states are more willing to overrule local leaders to prevent interference with ICE. But in both Democratic-controlled and Republican states, police have cooperated actively with ICE, and tend to provide basic security for ICE officers. Actual government intervention to prevent ICE enforcement has been lackluster. Instead, the most effective interventions have been through civil society organizations providing shelter and legal aid.

In this context, I would lean towards the idea that civil society mobilization is correlated with *both* liberal governors *and* lower ICE arrest rates and. that the correlation is mostly spurious.