Instructions

1. Don't panic. Historically, it takes real work to get worse than a B+. And, while grades do matter, getting a B+ or A- isn't going to change anything in your life. Take hard courses and shoot the moon once in a while (and make sure you take enough easier classes to maintain a GPA high enough to get past the filters – around a 3.4). Also remember to be clear / sharp in what you're saying so I can follow your thoughts; copying stuff from the internet you don't really understand won't help.
2. This is open book / open note / open internet. You **cannot** talk to other people about the exam.
3. I must have the exam by **Dec. 18** at 2 PM. Send your answers to our **private** slack channel as a stand-alone file – no links to cloud storage. And you must get a confirmation from me that I have it. If you don't hear back in a few hours, call me to make certain.
4. You may take the exam anywhere. But, leave yourself enough time to account for power outages, angry yeti, etc.

Questions (answer all 3)

1. Given the bias-variance tradeoff, how do you evaluate the role using a PCA has in selecting features for your regression model? To be more precise, imagine you have 10 independent variables and a two dimensional latent space captures 70% of the total variance. Will using these latent variables improve your regression model's ability to generalize out-of-sample? Or are there possible downsides? Which variables would you include in the PCA and which would you include separately in the regression? Finally, if you are running a purely predictive model, why or why not would you use feature selection of this kind?
2. Imagine you are trying to run a campaign for a presidential candidate. In the US, these campaigns compete in each state and it is winner-take-all for the candidate that wins each state. If you wanted to build a predictive model of how each candidate would do and identify factors that would help you advise your candidate, what unit of analysis would you focus on (i.e., what does a row in your dataset look like)? What challenges to inference exist, especially with respect to strategic behavior? What IVs would you collect?
3. For a given sample, you start with a purely linear regression model and then try a polynomial regression of order 3. In both cases, model fit is similar using MSE and both models show heteroskedasticity. Which models would you prefer, all else equal? Given the presence of heteroskedastic errors, what do you assume is wrong with your approach? How can you correct this problem? Finally, let's say you try a decision tree approach and MSE improves dramatically. What would you infer about the relationship of $y \sim f(\mathbf{x})$?

Data problem (mandatory)

With the attached data, your DV is modern day inequality (measured by gini_disp; see https://en.wikipedia.org/wiki/Gini_coefficient) and your IV's are various measures of countries at different points in time. Your sample is small b/c there are only so many countries in the world. Turn in your "best" model and a brief explanation of why you did what you did. Variables are as follows:

Ygini_disp	DV on inequality
country	Country name
federalism_GT	federalism variable
id	Country ID
region_wb	regional dummy
gdp	gdp
statehiste1500_02n	state agricultural history at 1500 AD
origtime2	origin time of state
eleva	elevation
avg_temp	average temp
Maddison_gdppc_1990_estimate_ln	gdp / capita in 1990
lp_lat_abst_fill	latitude
mountains	
log_ocdistance_new	distance from center of country to ocean
rugged	
tropical	

pmean	preciptiation mean
irri_impact5	impact of irrigation
frstdays	frost days
sd_emeanclip	variance in elevation
Urbanpopulationoftotalpop	
dist2suitable_km_new	distance from center to port
Fixedtelephonesubscriptionsp	
Employmentinagricultureof	
Accesstoelectricityofpopu	
pln_sxHr_mean	plantation crop suitability
agyears_ext	length of time using advanced agriculture
popd_1500AD	population at 1500 AD