**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?

**Given the bias-variance tradeoff, how do you evaluate the role using a PCA has in selecting features for your regression model?:**

So first what is the bias- variance tradeoff:

Random error is made up from bias and irreducible error. Bias can be theoretically reduced so the random error is just the irreducible error. In other words *Error = bias + variance*. Holding the error fixed yeads to an inverse relationship between bias and variance. High variance could lead to overfitting the dataset and poor results on new data. A model with high variance won’t work well at representing a new dataset as it is likely overfit to the training data. High bias could show consistent results between datasets but underfits the data. A model with High bias would work consistently between datasets but would not produce a result accurate to the data it receives. A naïve model has high bias.

We want a model that has a balance between bias and variance, It should be able to work on other datasets and still be responsive to difference in the datasets.

PCA modifies the variables used in your dataset to help lower variance and combat overfitting. PCA reduces how many dimensions your model has and less dimensions make a model harder to overfit and easier to understand by humans. PCA also helps reduce the amount of colinear variables. We want our IVs to be independent from each other as we are trying to find the effects of each variable onto our y. If we determine a bunch of different variables are related to each other (colinear). Then we can all those IVs and combine them into a new PCA variable. This reduces the dimensions of our dataset from how many IVs we put into our PCA into how many PCA variables we end up using.

Diagram

Description automatically generated

Here we can see the use of PCA to reduce 3 dimensions into 2. A linear regression can easily be done in a 2d space but is much harder to do in a 3d space.

I see PCA as a very useful way to prevent overfitting and reduce variance. It helps to reduce dimensions and combine colinear variables.

Will using these latent variables improve your regression model’s ability to generalize out-of-sample?:

Yes, the main problem with a model not being able to generalize out-of-sample is overfitting. Overfitting is from high variance in the model . It happens when a model pays attention to small changes too much. An overfit model would work really well for the dataset it is trained on but not work well on any other dataset. A way to counteract overfitting is to use latent variables.

Or are there possible downsides?:

Which variables would you include in the PCA and which would you include separately in the regression?:

if you are running a purely predictive model, why or why not would you use feature selection of this kind?

1. 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?
2. 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 ~ *f*(**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)](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 |