Keep in mind that we will want some commentary on how well the models generalise to the test set though

Another question regarding the preprocessing section of the assignment, did you want us to comment on the quality of the data (e.g. in the tutes we checked for occurrences or 0's, NaNs and multicollinearity, etc

What are the "key parameters" for ordinary linear regression? Are they just R^2, adjusted R^2, and test set RMSE? Would it be useless to mention p-values since regularized regression doesn't have a summary for them?

I would R^2, RMSE, p-values, etc are all properties. Parameters are things you set, or learn.

the QQ-plot and histogram show the same thing, so certainly one of these could be dropped,

Would discussion into the data quality (eg 0's, nans, ect) and the way the splits have been divided go into the 1st (Pre-Processing) or 2nd section (Trained Model Discussion) of the report?

Could go either - probably noting this stuff in pre-processing, and then commenting on any issues that arise from this in the discussion, would be gold standard, but if the info's there we'll take it. We're not too fused on having things in strict section

It's up to you if you want to standardise.

The data is already normalised (albeit just with a min/max normalisation), so you can certainly make the case that there is less/no need to standardise.

n the evaluation section, it states "Discussion of accuracy/validity should also consider the socio-economic nature of this data, and if or how this impacts the required accuracy and/or utility of the resultant models." What does this mean and what kind of information is required to discuss. And how does this differ from the last part about the ethical concerns of the socio-economic data.

Really we just want you to think about what sort of accuracy would be needed for this sort of model. When you report things like RMSE and R^2 rather than just saying "values of X, Y and Z were obtained, which shows the model is awesome/shithouse" consider a bit more if this sort of level of accuracy is appropriate given that such a model could be used to make funding and/or policy decisions.

DATA ALREADY NORMALISED

For this assignment is it necessary to use .add\_constant

That depends on if you standardise the data.

If you don't standardise, then you'll have a non-zero constant which you'll need to learn when you fit the model.

COMPARE ALL MODELS WITH TEST SET

any regularised model should be at best no more accurate on the training set (and that would be with lambda = 0). Accuracy should be lower on the training st.

It is somewhat subjective, but generally when you've got data in different units and/or different ranges it's good.

You'd certainly expect to see worse RMSE on the training set becuase that's the whole idea of regularisation, trade off some accuracy on the training data for a simpler and hopefully better generalising mode.

Ideally performance improves on the testing set - but it may not. You should see better performance on the validation set though

hanks Simon. In that case should we only be explaining lambda for this dot point?

Are there any other parameters I should consider?

Screenshot 2024-03-23 at 2.29.33 pm.png

Lambda mostly.

Other details (like R^2, RMSE, etc) are more in the performance section. Similarly, p-values, etc, probably fit elsewehre.

"better" if often difficult to quantify, as it depends on what you want - maybe accuracy is all you care about, maybe you need a simple model, maybe not violating assumptions is the most important thing in your setting, maybe there's some other criteria.

All we want you to do is look at the various models and their various properties and comment on that. If they're all crap, tell us why; if they're all the same, tell us why. Just be sure to explain and justify what ever conclusion you reac

QUESTION 2

