**Draft Project Feedback**

**General**

This is a well-constructed piece of work, with a strong flow – it’s clear chapter to chapter how you are building on what came before, and why each aspect of the project is important to the whole. You cover multiple methods, and often give good insight into their value in the current context.

There are though some areas (particularly Chapters 4 and 5) where you run through some quite dense theory without sufficiently explaining what is happening at each point. This can give the unfortunate impression that you are simply summarising the underlying theory, rather than explaining it having understood it. Your use of worked examples at the end of chapters does help to address this somewhat (I should note here that I did not check each line of calculation), but there is ultimately a difference between demonstrating you can replicate a process, and that you understand its implications. Working on this element would be an excellent way to further elevate your project.

**Declaration**

Don’t forget to add this!

**Chapter 1**

This is a very concise description of what your project will consider. Usually, concision is a valuable quality, but I think actually your introduction is \*too\* concise, and would benefit from some explanation. There are several ways this could be done:

* An explanation of what OE and sustainability scores are – how they are found, how they can be interpreted;
* A broader description of your data set in general, and the value to the reader of considering it;
* A summary of what each of your chapters brings to the overall project, to give the reader a sense of the project’s structure and flow;
* A little mor on the Exams Scores data set, what you’ll be using it for, and why you chose it for that role.

That said, what’s here is good, I’d just like to see it built on.

P4 para 1 “Single-response regression models 1 response…”->”Single-response regression models one response…”  
“…what if we need to model more than 1 response?”->”…What if we need to model more than one response?”

**Chapter 2**

This is a solid chapter, describing well the process by which the data was made ready for analysis. My comments here are all pretty small. I might hold back on in including R code when discussing what you’ve done, as it doesn’t really add value to the reader, but that’s a very minor point.

One thing that I think would be worth adding to this section would be a brief look at the relationship between the two response variables, to give some context to what is coming. In particular, it’s worth making the point that there is **some** relationship between them, otherwise there’s no real reason to apply multiple response regression here – we could just build one model for each response and call it it a day.

P5 para 3 It would be worthwhile going into detail about what the category values for the “category” and “risk\_rating” values actually are. It would also be useful to discuss the five (presumably) ordinal categories for “rating” in more detail than an overall comment about them relating to a desire to buy or sell.

P6 para 4 It is too strong to say “roe” and “sustainability score” are normally distributed. Instead, I would say we have strong evidence that they are at least approximately normally distributed. I’d also say that this comes from the fact the points on the QQ plot approximate a straight line, as oppose to this being a feature of the data points themselves.

P7 para 1 “Although high correlation implies multicollinearity, the opposite is not necessarily true.” It might be worth briefly explaining why this is the case, as a way of demonstrating your understanding.   
“…so it is essential to consider both variable types in the analysis.” Is it? How does VIF work in the context of categorical variables? What meaning does multicollinearity have for such variables?  
para 3 “Rating and risk rating seem to have no association…” It wasn’t clear to me why you’ve specified those two variables in this comment.  
para 8 “0.0.0862” is presumably meant to be “0.0862”. Also, you don’t need the whole dataset to demonstrate multivariate normality, so it would be useful here to be more specific about what you’ve applied Mardia’s test to (and to give a reference to said test).

**Chapter 3**

This is a strong chapter, giving generally good explanations of important and relevant concepts. Applying these approaches to a small example at the end of the chapter is a sensible move.

One thing that I wasn’t able to understand though is how you’re combatting overfitting. That is, how are you avoiding a situation in which adding additional predictor variables always reduces the amount of variance not explained by the model? This is a common problem for single-response regression – the SSR goes up every time we add a variable. There doesn’t seem to be anything stopping an equivalent phenomenon for lambda here. You do at the bottom of page 16 that you might be evaluating the models you’re building via cross-validation, which would remove this issue, but it isn’t really clear. Making sure this is addressed should be a priority.

P8 para 2 I’d call **Y** a response variable vector, rather than just the response variable. Same for **epsilon**.  
Homoscedasticity means we expect to see an even spread of residual values, and that we can assume the model operates equally well for all values of predictor variable. It says nothing about variable correlation. Also, homoscedasticity does **not** prove normality, though it is a necessary condition for the assumption that \epsilon\_i~N(0,\sigma^2) to hold.

P9 para 1 “…including the column of ones.” You didn’t mention this in the one-response case – best to do so to be consistent (and possibly explain what role that column has).  
para 2-4 I think these paragraphs would benefit from a rewrite. I think you might be using observations in two separate ways here (a single data vector, and an individual value of a response variable within a trial), and you’ve not defined what a trial is. You talk about a covariance matrix with subscripts i and k, but you only define what those actually are the second time you discuss that matrix. This page is key to giving the reader full understanding of what your approach involves, so it’s extremely important everything is fully clear.  
para 6 “…can be determined only from the observations **Y**\_(i)…” It wasn’t clear to me whether you mean this is a general truth, or a result of the assumptions of the MRLR approach itself.

P10 para 3 “where the linearity property of covariance does the first expansion and see A.2.1 for the simplification of Cov(**XB**) to **B**^TCov(**X**)**B**.”->“where the linearity property of covariance justifies the first expansion (see A.2.1 for the simplification of Cov(**XB**) to **B**^TCov(**X**)**B**).”

P11 para 4 “…iteratively adds the most useful predictor…” how is useful being defined here?

P12 para 1 You’re not consistent here as to whether you’re only adding one predictor at a time, or potentially more than one. It would also be useful to note when in the project sequential ANOVA is going to be made use of.

P13 para 3 Wilks’ Lambda test needs a reference.  
para 5 You use p here both as the number of dependent variables, and as shorthand for a p-value.  
para 6 “…by the newly added or removed predictor.” As before, you seem to be jumping between only changing the model by one predictor at a time, and allowing for a change by multiple predictors.  
para 8 “…evaluates the model for said response variable…”Not quite sure what you mean here – perhaps “…the model for a single response variable at a time”, or similar?

**Chapter 4**

This is a busy chapter, with an awful lot of theory thrown at the reader. The effect can make parts of this tough to read, as idea is piled atop idea without always a clear sense of whether what’s happening is fully justified. You do very well to compensate for this with your thorough run-through of these methods in Section 4.2, but it still might be worth taking a look at Section 4.1 and seeing whether there is any additional context/justification you can add as you go through these approaches. With that done, I think this chapter will go from solid to excellent.

Another thought I had was whether the two models you create in this Chapter could be compared to the OLS approach itself – presumably this will do less well, and thereby justify the approaches you’ve taken here.

P18 para 2 “The previous methods…”->”The methods given in Chapter 3”, or similar: anything being referred to from a previous section (or often just subsection) should be explicitly identified.  
“…coefficients are minimised or zeroed” I would rewrite as “…coefficients are reduced in magnitude, and perhaps even set to zero”, or similar. I’d also replace “…general solution…” with something along the lines of “…general loss function…”. We’re not looking to solve model fit + penalty, we’re looking to minimise it.  
para 3 “Ridge regression zeroes estimated regression coefficients…” assuming I have correctly understood what you mean by “zeroes”, I don’t think this is true – you note yourself a few lines later that ridge doesn’t reduce coefficients to zero.  
para 4 “It is useful when predictors hae high multicollinearity and when p is close to n.” Why is this the case? Discussing this would be a good way to demonstrate understanding.

P19 para 2 “A geometric interpretation of this interpretation is a diamond…” This is only true in the 2D case. It might be worth discussing the 2D case more, in fact (perhaps with figures) to help give context to the reader, and demonstrate your own understanding.  
para 5 It’s not clear what w^2\_ij represents in your explanation of the Frobenius norm, nor why we’re using that norm more generally. Why does it replace the l\_2 norm?

P20 para 3 What is the benefit of forcing a low rank?  
para 5 What is an “augmented dataset”?  
para 6 You state that augmentation is more efficient than using equation 4.1, but why is this the case?

P21 para 7 You state here you are plugging terms into equation 4.2, but it isn’t clear to me how the equation given here relates to 4.2, beyond them both having the same left hand term. Could you give more justification here, perhaps?

P23 para 2 “4.4 and 4.5 both have distinct ways to be solved”. This is slightly odd English – perhaps something like “4.4 and 4.5 are each solved using a distinct approach.”  
point 1 It’s not at all clear (at least to me) what “…the minimiser of the objective function along its coordinate direction…” means.  
para 3 “4.5 is solved using the graphical lasso (glasso) algorithm because it is fast and the most commonly used method to solve it.” The reference to speed here is fine, but the rest of this is circular: we solve this problem this method because this is the method most often used to solve this problem. Or do you mean you’re using a common approach to solving this problem in general to your specific case? If so, that’s perfectly sensible, and it just needs to be a bit clearer that that’s what’s happening.

para 4 I would note explicitly that this is where Algorithm 1 begins.  
  
**Chapter 5**

As with Chapter 4, (though to a lesser extent) there’s quite a bit of theory thrown at the reader in a short space of time. This could do with expanding a bit with some commentary on what is being done at each point, and why.

P30 para 2 “The prior discrete and continuous methods…” Again, new chapter, so you need to give the names of what you’re referring to here (and preferably where they can be found in the project). I’m also not quite sure what you’re referring to as a discrete method.  
para 3 “…fits a simple constant in each subgroup.” I’m not sure what is meant here – do you mean the method predicts the same category for all data within a given subgroup?  
Figure 5.1 is a very nice way to show what’s going on.

P31 para 3 It would be worth explaining what “pruning” means in this context.  
para 5 What are B an b in the formula for \hat{f]\_bag?  
“It also addresses overfitting through pruning, which was explained previously”. As noted, I don’t think you **did** explain it.  
para 6 you need either “Bagging is advantageous…” or “…bagging Is a good option…” here; not both.  
Section 5.1.3 needs at least one reference.  
  
P32 para 4 What dos is mean to “extract” an oob sample?  
para 8 “To avoid over-fitting, tree splitting and optimal tree selection, use CV and pruning, like CARTs”. I’m not sure how to parse this sentence – do we want to avoid optimal tree selection? Or do you mean that, to avoid overfitting, we will use CV and pruning when considering tree splitting and optimal tree selection?  
  
P33 Figure 5.2 is a very nice addition, but at present it’s quite underexplained. What does the Y rectangle represent, or the rectangle broken into three subrectangles x1, x2, and x3? The right hand side of the picture is easier to understand, but I assume you mean “histogram” rather than “bar plot” (since both responses are continuous), and it’s not clear how one histogram will suffice considering we have a pair of response variables. It would also be worth noting that (assuming I’m reading the graph correctly) the nodes with fewer bars have that property because the distribution of the corresponding response variable is narrower.

**Chapter 6**

This reads well, and I feel offers a better explanation of the underlying theory than was present in Chapters 4 and 5, though there are still occasions where it isn’t fully clear what is being done and why – see below.

P39 Your short summary at the top of Section 6.1 would work better as the first paragraph to the whole chapter – it’s not clear from your current first paragraph what the chapter is really about.  
para 3 What is T here? I assume it’s the number of trees generated so far, and that the idea is we increase T, adding a new term to the series each time (should not the series be divided by T? Otherwise, won’t our estimate will just keep getting bigger?).   
para 4 What does phi here represent?  
para 5 So I’m guessing this T (for number of leaves) is different to the T (in a different font) in the two equations above it? Also, how are the weights defined/determined?

P40 para 2 What is eta here?  
para 6 How are left and right nodes defined? Is the implication here that the tree only ever performs binary splits (perhaps you’ve already noted this in a previous chapter and I’ve not noticed/remembered)? How is the pruning parameter gamma defined/determine?

P42 para 3 I’m not sure what the implications are of applying exponential functions to the variances – can this be done, and what does it mean for the underlying data? How does restricting the coefficient of correlation help?

**Chapter 7**

Again, there’s some hefty theory here, which would benefit a great deal from additional commentary to aid the reader (see below).

P48 para 1 “The final chapter of this report…” This isn’t the final chapter. You could say it’s the final theory-based chapter, though.  
“Ending with…”-> “We shall end with…” or similar.  
para 4 “The first of which…”->”The first of these…”  
  
P50 para 1 “The first of which…”->”The first of these…”  
What does “rushed” mean in the context of optimisation?  
para 6 I think some additional discussion of your decision to introduce task specific layers would be useful here – it isn’t clear to me how the (comparatively low) correlation between the two response variables means they shouldn’t have the exact same input nodes. Additionally, based on how you have described the nodal structure of these models (and remember I’m not at all an expert in neural nets), I don’t see how additional internal nodes can resolve an issue you’re claiming relates to the input nodes. It’s additionally unclear as to what elements of the work that follows in Chapter 7 involve your bespoke method. We’ve talked previously about the importance of fully justifying any decision to attempt your own modifications to established methodology, and this paragraph gives you a great opportunity to do just that.   
(Actually, does this get used at all after you mention it here? I’m not seeing where).

P51 I’m struggling to understand what our original points **X** and our new points **X\_\*** represent here, which makes the explanation overall difficult to follow.  
para 6 I don’t think you’ve proved the conditional distribution here – it doesn’t look unreasonable given your previous comments/expressions, but that’s very different.  
para 9 “owe”->”we”. Also, its not clear what a radial basis function is, or how it helps address the problem you have identified.

P55 para 2 “…which is the best model.” I assume you mean here the best model so far?

**Chapter 8**

A sensible way to end the project. One way this could be expanded would be to talk about what the various models (especially those that worked particularly well) actually tell us about the underlying structure of the data – you note exploring this as a goal early on, but it doesn’t really come through as a focus.  
  
P57 5.1.2 I’d remove the numbers here – they’re available in Table 8.1 just over the page, and replicating them here interrupts the flow of what you’re saying.  
para 6 “The first of which…”->”The first of these…”  
It's common to include a discussion of data limitations in a section on challenges. Was there anything about the dataset you created that could have been improved given more time or with the benefit of experience, or any aspects of it that caused particular challenges?  
  
**Bibliography**

There’s a smattering of missing capital letters here – remember, journal titles and book titles are capitalised in a bibliography, while individual paper titles are not.