Assignment 2
Data Science Methods:

André Couder (2070121) Daniel Redel (2102630)
Frederico Sandmann (2071361) Xander Smid (2093581)

March 21, 2023

Question 1

Question a)

For starters, we used a sample size of 1000 due to computation constraints, but kept all the variables. We used 800 observations for the training set and 200 observations for the test, as is often used in practical application. Finally, for the k-fold cross validation approach we used a $k = 5$.

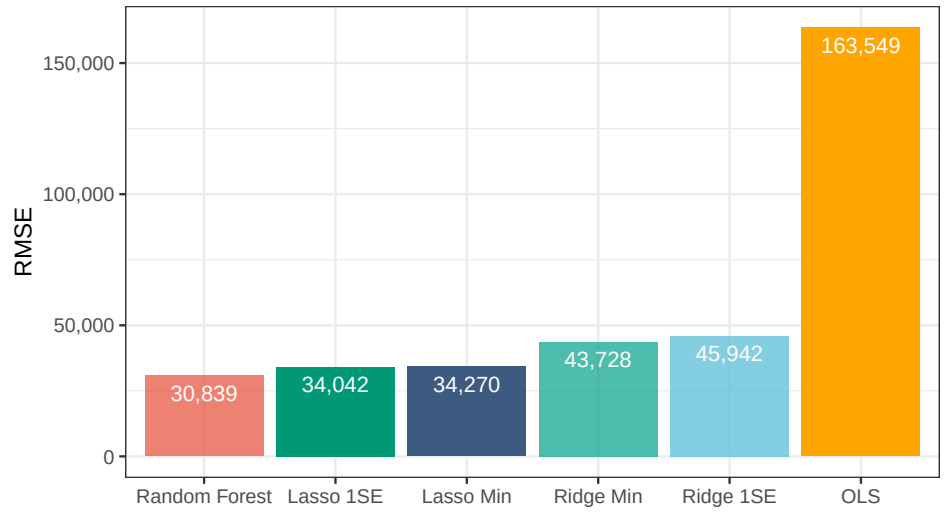

Figure 1: Model Comparison: Out-of-Sample RMSE

As can be seen in Figure 1, lasso, ridge, and random forest all out-perform the linear model in terms of the out-of-sample MSE. For completeness, besides adding the lasso and ridge with the lambda that minimizes the CV MSE, we also included the lasso and ridge with the tuning parameter λ that yields one standard deviation above the minimum CV MSE (rule of thumb that is often used). As can be seen in Figure 1 (using the simple validation approach) ***random forest is the method that has the smallest out-of-sample MSE out of all the methods***, given our sample and split.

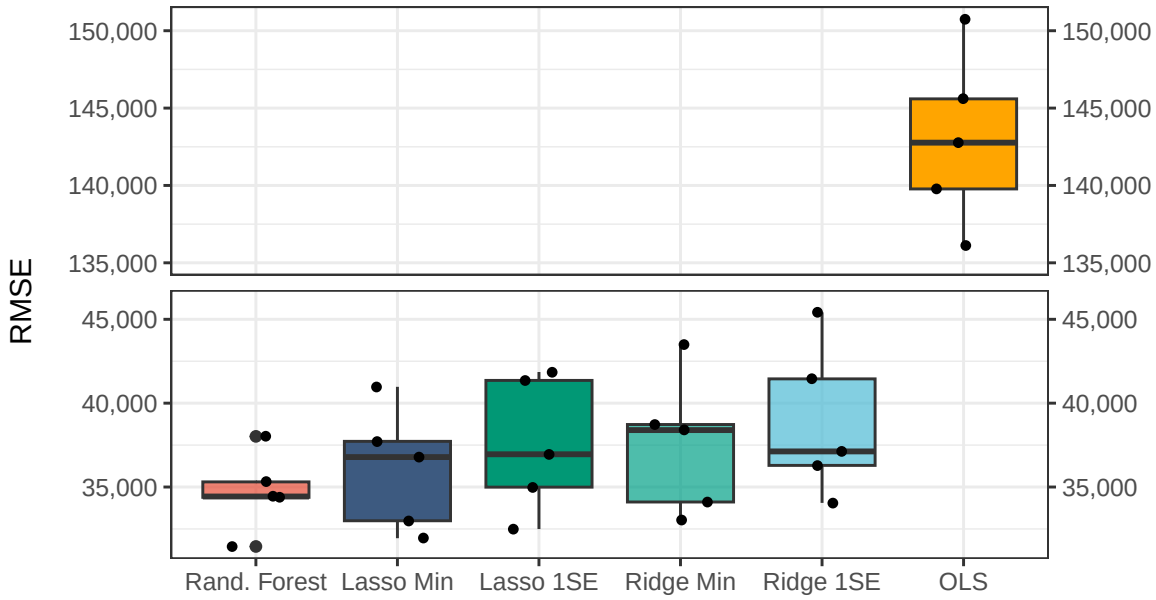

Figure 2: Model Comparison: k-fold Out-of-Sample RMSE

This is also the case when looking at Figure 2 (using k-fold cross validation), however, the large confidence intervals indicate that the difference between the approaches might not be too large. This leads us to be believe that we should be careful about giving definitive answers about the best model in this finite sample case. Still, **from now on we will consider random forest as the best approach**. Normally, random forest tends to work better than other approaches when the model is complex (eg. highly non-linear, with complex interactions between the variables), and this is the case in our Figure 1, implying that our model might be quite complex. Ridge and lasso, on the other hand, work best in linear models. Note also that although Random Forest seems to perform better in 5-fold cross validation, The confidence intervals in Figure 2 are substantial, implying that one should be careful about taking definitive conclusions about the superiority of a method, even so we will work with random forest from now on(regarding the 3 close methods). As expected, random forest also has a lower test MSE

than OLS due to the large number of variables and possible complexity of the model.

All in all, it seems like all methods outperform OLS substantially (in terms of out-of-sample MSE) but this difference is much smaller between the remaining approaches (especially in Figure 2). ***We will consider random forest as the best method*** and, since in the labs we used the simple validation method, we will focus on it in `b)`.

Question b)

In order to highlight the best covariates in terms of prediction, we considered different measures of variable importance, as we will describe below. Figure 3 reports the most relevant variables in each method:

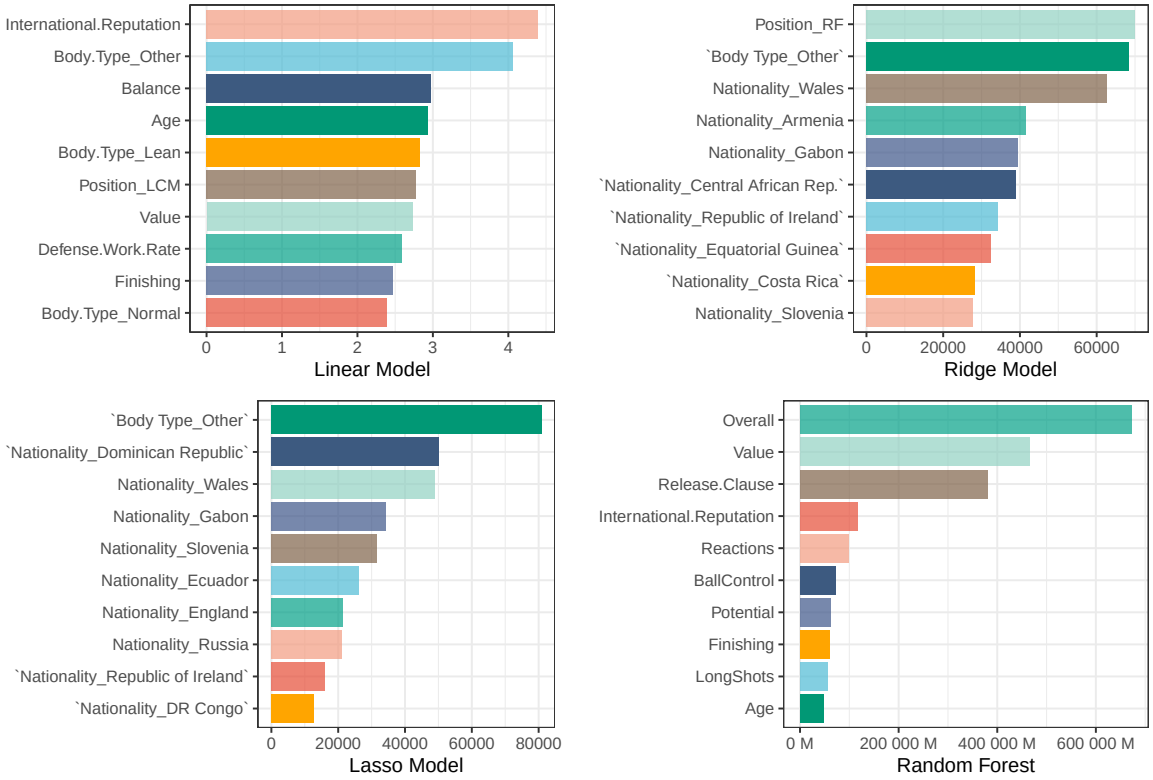

Figure 3: Model Comparison: Variable Importance

Linear Model:

- First of all, note that we have standardized our data in this whole exercise. This preliminary step ensures that if $|\hat{\beta}_i| \geq |\hat{\beta}_j|$ for some covariates i, j , the model returns a relative greater explanatory power of covariate x_i compared to covariate x_j regarding

the training dependent variable (provided that both standard deviations i.e. of $\hat{\beta}_i$ and $\hat{\beta}_j$ are equal). This feature led us to consider as best predictors the covariates with biggest $|t_\alpha|$ (absolute value of t-statistic).

- The chosen covariates are the following: `International reputation`, `Body type: Others`, `Balance`, `Age`, `Body type: Lean`, `Position: LCM`, `Value` , `Defense work rate`, `Finishing`, `Body type: Normal`.

LASSO Model:

- For LASSO Model, we considered 2 ways to pick the 10 most important regressors (we wanted 10 to compare the most important regressors with our other models).
- First of all, we considered the variables corresponding to the **biggest absolute z-scores of parameters**, among the 26 non-zero parameters, returned by the model for the optimal λ . We used the z-scores to take into account the correlation between variables that may impact the relative importance of each variable with respect to the 25 others for prediction. These outputted variables can be seen as the 10 most important out of 26 with a non-zero coefficient in the model. Note that, of course, if we have 26 variables with a non-zero coefficient returned by the model, we gain in predicting power by keeping them all.
- The most relevant variables given this method are: `Body type: Others` , `Nationality: Dominican Republic` , `Nationality: Whales`, `Gabon`, `Nationality: Slovenia`, `Nationality: Equador`, `Nationality: England`, `Nationality: Russia`, `Nationality: Ireland`, `Nationality: DR Congo`.
- Second, and as an alternative measure, we took the value of λ that shrinks all coefficients to 0 except for 10. This outputs the 10 most important variables for prediction **given that every other coefficient is set to 0** such that correlation is not an issue. These results are shown in the right-hand of Figure 4 below.
- The most relevant variables given this method are: `Value`, `Overall`, `International Reputation`, `Release Clause`, `Body type: Others`, `Nationality: England`, `Sliding tacle`, `Nationality: Whales`, `Age`, `Defense work rate`.

Ridge Model:

- To extract the 10 "best" predictors from the Ridge regression, we used the **ridge trace method**. This method calculates the sum of the absolute values of the standardized coefficients (of Ridge regressions) across all values of λ in the ridge path (the successive ridge regressions for each value of λ). This method has the advantage to take into account the contributions of all the variables to the model, across all values of λ. It therefore takes into account that some variables may be more important than others at different values of λ (because of the correlation between regressors coupled to different scalings for different λ).

- As an alternative method, we also checked which regressors were corresponding to the highest parameters in absolute value for the optimal λ (see left-hand of Figure 4). From Figure 4 we conclude that **both methods returned the same ordered list of variable importance**. This tells us that the most important variables in average over all values of λ are the same as the most important variables for λ_{opt}.
- The chosen predictors are listed below: `Position: RF`, `Body type: Others`, `Nationality: Whales`, `Nationality: Armenia`, `Nationality: Gabon`, `Nationality: Central African Republic`, `Nationality: Republic of Ireland`, `Nationality: Equatorial Guinea`, `Nationality: Costa Rica`, `Nationality: Slovenia`.

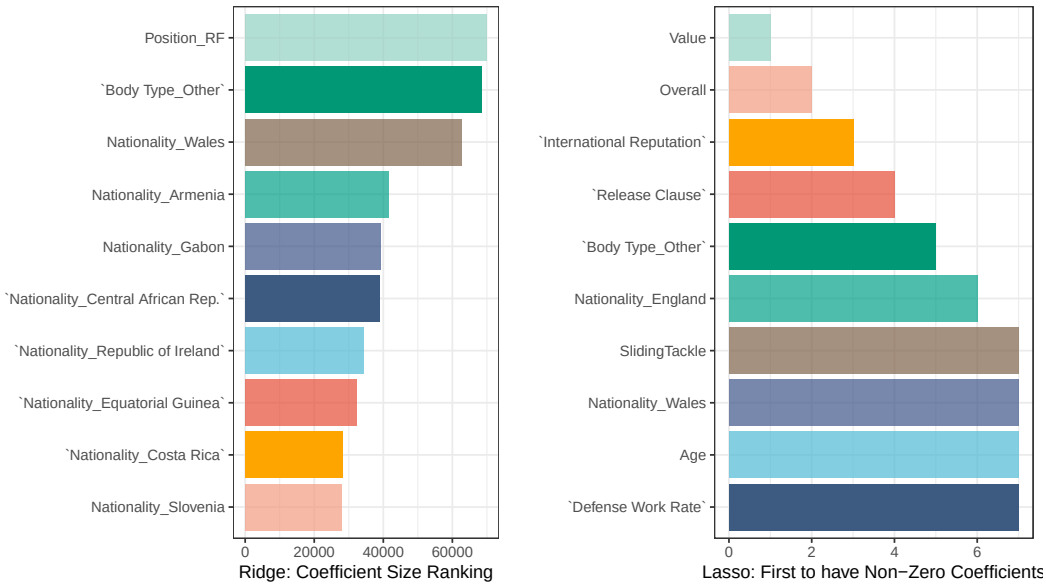

Figure 4: Ridge vs Lasso: Alternative Measures

Random Forest Model:

- The relative importance of predictor x_i in the `RandomForest` package is based on the *sum of the squared* improvements over all internal nodes of the tree for which x_i was chosen as the partitioning variable (Breiman, Friedman, and Charles J. Stone; 1984).
- In particular, to calculate the variable importance in the Random Forest model, we trained the model (grew N trees with a subset of the regressors) on the training dataset D and measured its accuracy A_1 (average RSS accuracy of the N trees) on the test set T_1, then we permuted each value of the variable x_i randomly in the test dataset, call this dataset T_2. Then we used the trained model to make predictions on this dataset T_2 and measured the prediction's accuracy, call it A_2. The normalized difference between A_1

and A_2 is our measurement of variable importance for x_i. This process was reproduced for each variable.

- The 10 best variables as shown in Figure 3 are listed below: `Overall`, `Value`, `Release clause`, `Intrenational reputation`, `Reaction`, `Ball control`, `Potential`, `Finishing`, `Long shot`, `Age`.

Question c)

We picked 3 players, namely: Cavani, Quaresma and Van Dijk. The model we used to predict the wages was **random forest** since that was the best performing model in `a)`. Figure 5 reports our predictions against the actual wages of each player:

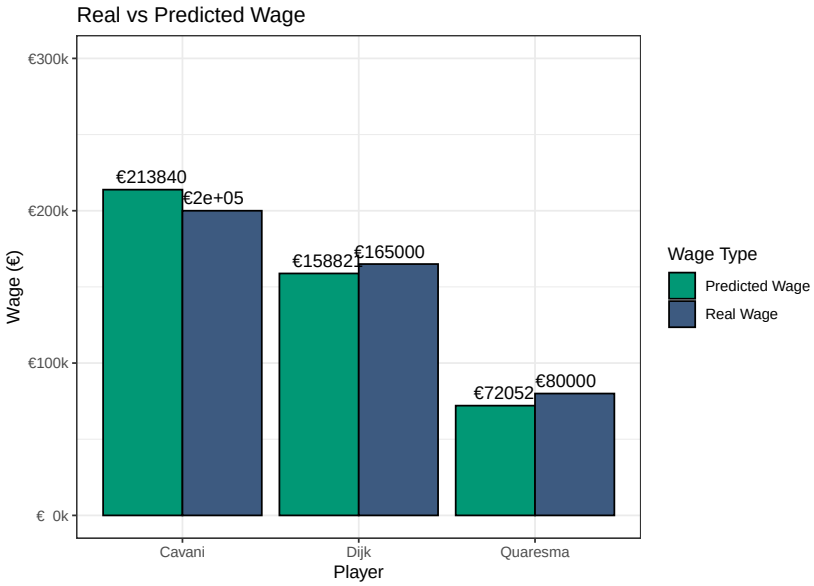

Figure 5: Players Predicted Wages

As we can observe from Figure 5, a player with the characteristics of Cavani would be expected to earn a higher salary than Cavani (almost €14k more). A player with the characteristics of Van Dijk would be expected to earn a lower salary than Van Dijk (about €7k less). Finally, a player with the characteristics of Quaresma would be expected to earn a lower salary than Quaresma (around €8k less).

Question 2

We consider the following model:

$$y_i = \beta_i + \varepsilon_i$$

where $\varepsilon_i \sim \text{iid}(0, 1)$.

Question a)

The penalized least-squares objective function is defined by:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \left[\sum_{i=1}^{N} \left(y_i - \beta_i\right)^2 \; + \; \lambda \sum_{i=3}^{N} \left((\beta_i - \beta_{i-1}) - (\beta_{i-1} - \beta_{i-2}) \right)^2 \right]$$

Question b)

We can re-write the previous objective function in matrix form:

$$\begin{aligned}
\hat{\beta} &= \underset{\beta}{\text{argmin}} \left[\sum_{i=1}^{N} \left(y_i^2 - 2y_i\beta_i + \beta_i^2\right) \; + \; \lambda \sum_{i=3}^{N} \left(\beta_i^2 + 4\beta_{i-1}^2 + \beta_{i-2}^2 - 4\beta_i\beta_{i-1} - 4\beta_{i-1}\beta_{i-2} + 2\beta_i\beta_{i-2}\right) \right] \\
&= \underset{\beta}{\text{argmin}} \left[y'y - 2y'\beta + \beta'\beta + \lambda \sum_{i=3}^{N} \begin{bmatrix} \beta_{i-2} & \beta_{i-1} & \beta_i \end{bmatrix} \begin{bmatrix} 1 & -4 & 2 \\ 0 & 4 & -4 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{i-2} \\ \beta_{i-1} \\ \beta_i \end{bmatrix} \right] \\
&= \underset{\beta}{\text{argmin}} \left[y'y - 2y'\beta + \beta'\beta + \lambda \beta' \, A \, \beta \right]
\end{aligned}$$

$$A = \begin{bmatrix}
1 & -4 & 2 & 0 & 0 & 0 & ... & 0 & 0 & 0 \\
0 & 5 & -8 & 2 & 0 & 0 & ... & 0 & 0 & 0 \\
0 & 0 & 6 & -8 & 2 & 0 & ... & 0 & 0 & 0 \\
0 & 0 & 0 & 6 & -8 & 2 & ... & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & ... & 6 & -8 & 2 & 0 & 0 \\
0 & 0 & 0 & 0 & ... & 0 & 6 & -8 & 2 & 0 \\
0 & 0 & 0 & 0 & ... & 0 & 0 & 6 & -8 & 2 \\
0 & 0 & 0 & 0 & ... & 0 & 0 & 0 & 5 & -4 \\
0 & 0 & 0 & 0 & ... & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

Question c)

We want to find the β that minimizes the in-brackets expression above. Note that this expression is quadratic in β and that:

- Given the algebraic form of the constraint above, i.e. $\lambda \sum_{i=3}^{N} \left[(\beta_i - \beta_{i-1}) - (\beta_{i-1} - \beta_{i-2})\right]^2$ we have that the quadratic expression$\beta' A \beta \geq 0$ thus the matrix A is positive-definite and $\lambda \beta' A \beta$ is convex if $\lambda \geq 0$
- Given the matrix form of the equation above we have $\beta'\beta = \|\beta\|^2$ is convex.

Therefore we can find $\hat{\beta}$ by setting the derivative of the inside-brackets expression above to 0 .

Let be Obj $= y'y - 2y'\beta + \beta'\beta + \lambda\beta' A \beta$, then:

$$\begin{aligned} \frac{\partial}{\partial \beta}\text{Obj} &= 0 \\ -2y + 2\beta + 2\lambda A\beta &= 0 \\ (I_N + \lambda A)\,\beta &= y \\ \hat{\beta} &= (I_N + \lambda A)^{-1} y \end{aligned}$$

We can rewrite this closed form solution as $\hat{\beta} = (I_N' I_N + \lambda A)^{-1} I_N' y$ and we see that this expression resembles closely the **Ridge Estimator** assuming our data matrix X is the identity matrix I and furthermore noting that λ is multiplied by A instead of I because the constraint is different.

Question d)

1. **Expectation**:

$$\begin{aligned} \mathbb{E}\left[\hat{\beta}\right] &= \mathbb{E}\left[(I + \lambda A)^{-1} y\right] \\ &= \mathbb{E}\left[(I + \lambda A)^{-1} (\beta + \epsilon)\right] \\ &= (I + \lambda A)^{-1}\, \mathbb{E}\left[\beta\right] \\ &= (I + \lambda A)^{-1} \beta \end{aligned}$$

2. **Variance**:

$$
\begin{aligned}
\operatorname{Var}\left(\hat{\beta}\right) &= \mathbb{E}\left[\hat{\beta}\hat{\beta}'\right] - \mathbb{E}\left[\hat{\beta}\right]\mathbb{E}\left[\hat{\beta}'\right] \\
&= (I+\lambda A)^{-1}\,\mathbb{E}\left[yy'\right]\left((I+\lambda A)^{-1}\right)' - (I+\lambda A)^{-1}\,\mathbb{E}\left[y\right]\mathbb{E}\left[y'\right]\left((I+\lambda A)^{-1}\right)' \\
&= (I+\lambda A)^{-1}\,\mathbb{E}\left[(\beta+\epsilon)(\beta+\epsilon)'\right]\left((I+\lambda A)^{-1}\right)' - (I+\lambda A)^{-1}\,\mathbb{E}\left[\beta+\epsilon\right]\mathbb{E}\left[\beta'+\epsilon'\right]\left((I+\lambda A)^{-1}\right)' \\
&= (I+\lambda A)^{-1}\,\mathbb{E}\left[\beta\beta'\right]\left((I+\lambda A)^{-1}\right)' + (I+\lambda A)^{-1}\,\mathbb{E}\left[\beta\epsilon'\right]\left((I+\lambda A)^{-1}\right)' \\
&+ (I+\lambda A)^{-1}\,\mathbb{E}\left[\epsilon\beta'\right]\left((I+\lambda A)^{-1}\right)' + (I+\lambda A)^{-1}\,\mathbb{E}\left[\epsilon\epsilon'\right]\left((I+\lambda A)^{-1}\right)' \\
&- (I+\lambda A)^{-1}\,\mathbb{E}\left[\beta\right]\mathbb{E}\left[\beta'\right]\left((I+\lambda A)^{-1}\right)' - (I+\lambda A)^{-1}\,\mathbb{E}\left[\beta\right]\mathbb{E}\left[\epsilon'\right]\left((I+\lambda A)^{-1}\right)' \\
&- (I+\lambda A)^{-1}\,\mathbb{E}\left[\epsilon\right]\mathbb{E}\left[\beta'\right]\left((I+\lambda A)^{-1}\right)' - (I+\lambda A)^{-1}\,\mathbb{E}\left[\epsilon\right]\mathbb{E}\left[\epsilon'\right]\left((I+\lambda A)^{-1}\right)' \\
&= (I+\lambda A)^{-1}\,\beta\beta'\left((I+\lambda A)^{-1}\right)' + (I+\lambda A)^{-1}\,\beta\mathbb{E}\left[\epsilon'\right]\left((I+\lambda A)^{-1}\right)' \\
&+ (I+\lambda A)^{-1}\,\mathbb{E}\left[\epsilon\right]\beta'\left((I+\lambda A)^{-1}\right)' + (I+\lambda A)^{-1}\,I\left((I+\lambda A)^{-1}\right)' - (I+\lambda A)^{-1}\,\beta\beta'\left((I+\lambda A)^{-1}\right)' \\
&= (I+\lambda A)^{-1}\left((I+\lambda A)^{-1}\right)'
\end{aligned}
$$

Question e)

If there was no restriction in this problem we would find $\hat{\beta} = y$, and therefore:

$$\mathbb{E}\left[\hat{\beta}\right] = \mathbb{E}\left[y\right] = \mathbb{E}\left[\beta+\epsilon\right] = \mathbb{E}\left[\beta\right] = \beta$$

$$\operatorname{Var}\left(\hat{\beta}\right) = \operatorname{Var}\left(y\right) = \mathbb{E}\left[yy'\right] - \mathbb{E}\left[y\right]\mathbb{E}\left[y'\right] = \mathbb{E}\left[\beta\beta'\right] + \mathbb{E}\left[\epsilon\epsilon'\right] - \mathbb{E}\left[\beta\right]\mathbb{E}\left[\beta'\right] = I$$

Instead we have, as shown above:

$$\mathbb{E}\left[\hat{\beta}\right] = (I+\lambda A)^{-1}\,\beta$$

$$\operatorname{Var}\left(\hat{\beta}\right) = (I+\lambda A)^{-1}\left((I+\lambda A)^{-1}\right)'$$

Note that since the constraint can be rewritten as a sum of squares, we have that A is positive definite. Therefore we can rewrite by eigen-decomposition:

$$
\begin{aligned}
(I+\lambda A) &= (UIU' + \lambda U\epsilon U') \\
&= U\left(I+\lambda\epsilon\right)U' \\
(I+\lambda A)^{-1} &= U\left(I+\lambda\epsilon\right)^{-1}U' \\
&= U \underbrace{\begin{bmatrix} \frac{1}{1+\lambda\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{1+\lambda\sigma_2^2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{1+\lambda\sigma_N^2} \end{bmatrix}}_{V} U' \\
(I+\lambda A)^{-1}\left((I+\lambda A)^{-1}\right)' &= UV^2U'
\end{aligned}
$$

Note that given the above, we have :

$$\text{Var}\left(\widehat{\beta}_i\right) = v_1^2u_{1i}^2 + v_2^2u_{2i}^2 + \cdots + v_N^2v_{N_i}^2$$

where $v_i^2 = \frac{1}{(1+\lambda\sigma^2)^2}$ and $\sum_{j=1}^{N} u_{ji}^2 = 1 \ \forall\, i$ therefore we can only obtain $\text{Var}\left(\widehat{\beta}_i\right) = 1$ if $\lambda = 0$.

Furthermore, it follows directly from $\mathbb{E}\left[\widehat{\beta}\right] = (I+\lambda A)^{-1}\beta$ that $\mathbb{E}\left[\widehat{\beta}\right] = \beta$ if and only if $\lambda = 0$. thus the only possible lambda that mimics the absence of constraint is $\lambda = 0$ (i.e. no constraint).