1. i. The curse of dimensionality refers to the inherent dint in the performance of a model that trains and tries to predict data sourced from a high dimensional space. Although increasing the number of predictors i.e. features from zero would initially improve prediction performance due to having more features to input useful information, the data points become too sparse after including too many features (known as the peaking phenomenon). Parametric models are relatively less affected than nonparametric counterparts due to inherent interpretability from a fixed set of parameters formulating a relation between predictors; for example, a decision tree has exact boundaries of splits per feature. However, nonparametric models such as KNN would have to face a very large space with sparse data points when many features are used.

   ii. Logistic regression, LDA, and linear SVC are all linear classifiers but have their own nuances, weaknesses, and different coefficients for the linear form. Logistic regression applies the logit link in a generalized linear model to transform ordinary linear regression into binary classification problem; the class assignment depends on whether the logistic function value (must be in (0, 1)) is greater than a chosen threshold or not. Rather than a single linear function, LDA produces a set of k discriminant functions, one for each of k classes, and classifies a specific observation by choosing the class yielding the highest value from its discriminant function. The equation defining the dividing hyperplane for linear SVM only depends on the data points on or inside the margins, so the coefficients are very susceptible to variance; changes in data such as the case in k-fold CV will likely yield different coefficients for the hyperplane.

   iii. Yes, overfitting does seem more likely to occur than not. Recall that the RBF kernel function is $K(X, X') = e^{-\gamma\|X-X'\|}$. As γ increases, data points farther away matter less and less in contributing to defining decision boundaries. In other words, the radial SVM becomes more localized for the observations that contribute to decision boundaries.

   iv. Yes. Ordinary linear regression without the penalty term does fine, but adding a penalty term adds a constant that does not preserve transformations for matrices.

   v. Yes, some situations may render LASSO less capable than a stepwise or subset selection method. Suppose there are a few predictors that correlate extremely closely with the output Y but not so much with each other, and suppose all other predictors left in the model barely correlates with Y but correlates extremely high with each other. The first set of predictors mentioned would most likely all get swatted (set to zero) by LASSO as this type of regression pushes coefficients to zero when correlation with Y is high enough. This leaves a set of nonzero but intercorrelated and highly useless set of predictors left.

   vi. Focusing on $P(|Y - \hat{Y}| > 2)$ as the evaluation metric calls for a model that is tunable to allow some amount of misclassification to occur as long as the absolute error is less than 2 units. SVM, being a broader generalization of the optimal hyperplane to allow nonlinear decision boundaries, seems to be a good choice for the model. SVM has a parameter C that specifies the weighted count of misclassifications allowed across these boundaries. Depending on how much overlap the classes have, C can be tuned accordingly with a gridsearch strategy. Note that this is just one example, as

any other model with a tunable parameter ignoring penalties for misclassifications within some bound can be compared to SVM as well to choose the better model.

2. i.

$Var(\bar{X}) = Var\left(\frac{1}{n}\Sigma_{i=1}^{n}X_i\right) = \frac{1}{n^2}Var(X_1 + X_2 + \cdots + X_n) = \frac{1}{n^2}\Sigma_{i=1}^{n}Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

ii.

$E\left[X_i^{(b)}|X\right] = E\left[X_i^{(b)}\middle|X_1, X_2, \dots, X_n\right] = E\left[\frac{1}{n}\Sigma_{i=1}^{n}X_i^{(b)}\right] = \frac{1}{n}E\left[X_1^{(b)} + X_2^{(b)} + \cdots + X_n^{(b)}\right] =$
$\frac{1}{n}\Sigma_{i=1}^{n}E\left[X_i^{(b)}\right] = \frac{1}{n} * n\bar{X} = \bar{X}$

iii.

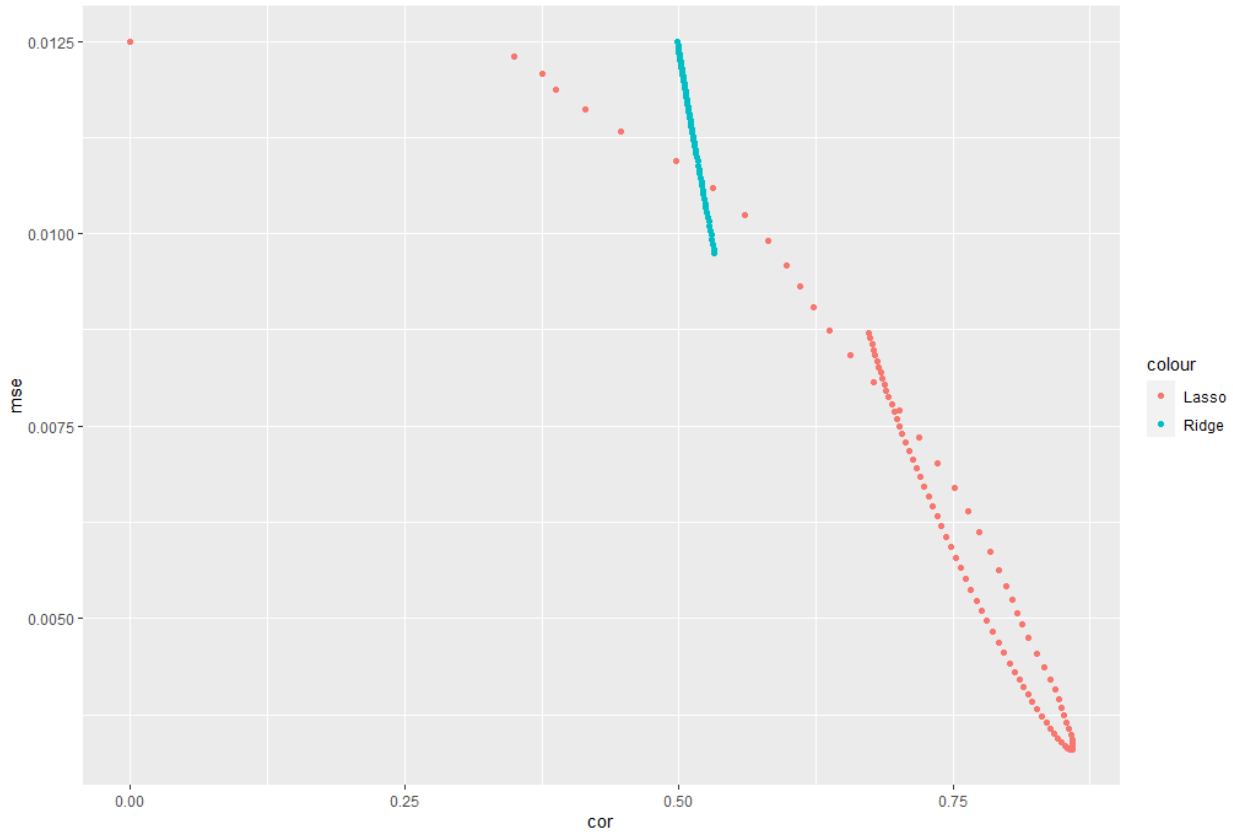
$Var\left(X^{(b)}\middle|X\right) = Var\left(X^{(b)}\middle|X_1, X_2, \dots, X_n\right) = \frac{1}{n}\Sigma_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$
$Var\left(\bar{X}^{(b)}\middle|X\right) = Var\left(\frac{1}{n}\Sigma_{i=1}^{n}X_i^{(b)}\middle|X\right) = \frac{1}{n^2}\Sigma_{i=1}^{n}Var\left(X_i^{(b)}\middle|X\right) = \frac{1}{n^2} * nVar\left(X^{(b)}\middle|X\right) = \frac{1}{n^2} * n *$
$\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n^2}S^2$

iv.

$E_X\left[Var\left(\bar{X}^{(b)}\middle|X\right)\right] = E_X\left[\frac{n-1}{n^2}S^2\right] = \frac{n-1}{n^2}E_X[S^2] = \frac{n-1}{n^2}\sigma^2$

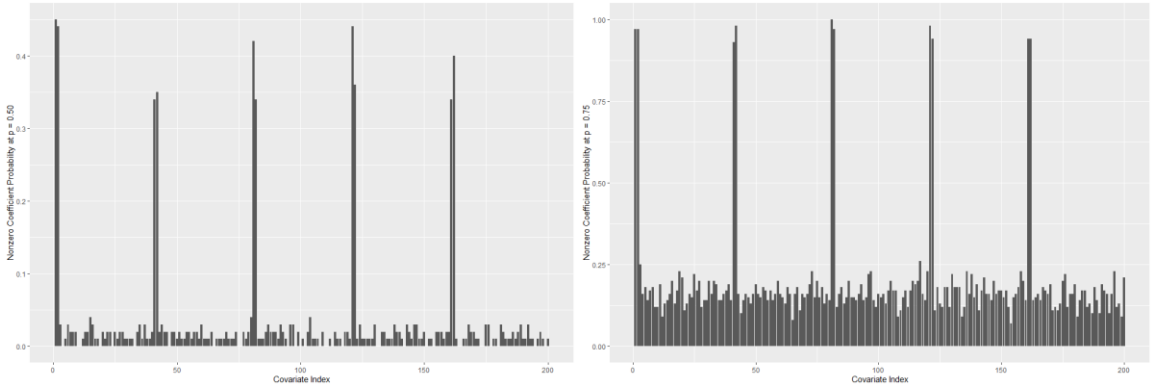
v. The results from parts i. and iv. are not the exact same but still extremely close when comparing the coefficients multiplied to σ^2. Although not theoretically equivalent in formula, bootstrap sampling does an accurate job of estimating the uncertainty of $\bar{X}$.

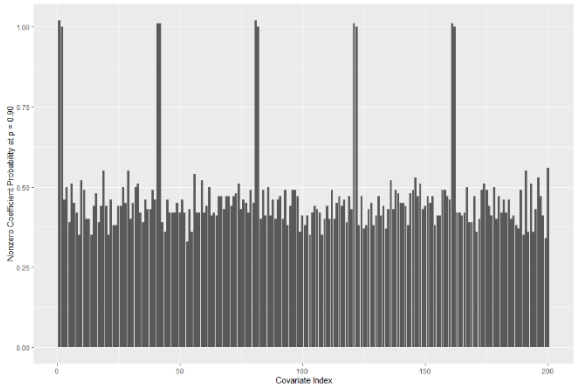
3. i. See code also.

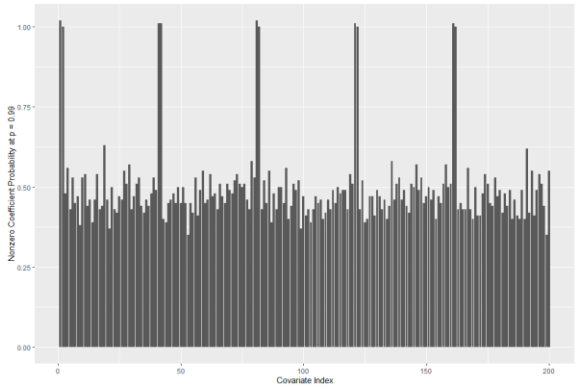


ii. As correlation increases, the MSE, averaged across 100 simulations, decreases as expected for both ridge and LASSO. However, ridge seems to decrease MSE at a steeper rate. Additionally, the lowest MSE that ridge achieved could not reach the lowest MSE that LASSO achieved. LASSO can remove predictors entirely while ridge just shrinks coefficients towards zero, which somewhat explains this phenomenon.

iii. Corresponding order of p=0.50, 0.75, 0.90, 0.99.







iv. Regardless of the correlation coefficient, the few sparse and nonzero coefficients originally mentioned in the question have high probability of being nonzero from the simulation study. As correlation increases however, the covariates known to be zero have increasing chances of being being chosen. As correlation increases, the relevant covariates, which presumably correlate better with Y, get swatted, therefore leaving more irrelevant covariates being used more often in model.

v. There definitely can be. One can try Elastic Net, which pushes coefficients towards zero like ridge but does not delete coefficients for predictors that highly correlate with Y, unlike LASSO.

4. i. MSE is 0.7112019. See code for predictions.

ii. The CI was constructed via bootstrapping on the training set, and then applying the percentile method on the parametrically computed values for $\beta_1 + e^{\beta_2} + \beta_3^2$ from the bootstrap samples. The work was carried out by the *boot* package. Please see the code for details. The CI was (1.288, 1.898).

iii. Nope. Redoing the same bootstrapping as a repeated experiment would almost always yield a different CI due to randomness of sampling. A different CI generated was (1.392, 1.982).

iv. Nope. The principal components derived by PCA only draws contribution from the predictors, but they have no association or correlation with Y.

v. The AIC for both the full model and the forward stepwise reduced model are nearly the same. There does not seem to be any significant improvement.

vi. There are numerous ways to check for the potential use of interaction terms. A simple preliminary measure is checking for correlation of the predictors amongst themselves. Predictors with high correlation e.g. > 0.80 can potentially be combined. Since high correlation between predictors would inflate their variance in the model, applying VIF can help pinpoint specific predictors to be combined. A VIF > 10 would be considered extremely high, which is the case for all predictors except X1 and X23.

vii.

Model	Error Rate
LDA	0.3089214
QDA	0.3089214
SVM radial	0.2163872
SVM poly	0.1948346