Assignment 1

Zhengyang Fei

Disclaimer: My report does not follow the format of Intro, Methods, Results, Conclusions that is described in the syllabus as stated in the assignment description. In my report and code, I followed the tasks one step at a time and provided step by step commentary, making clear what the final answer is and explaining graphs/tables.

Task 1

First, we import the dataset after setting work directory. Then we call read.csv() to get the dataset. The argument header is set to FALSE to tell R that the first row is not the column names. We proceed by properly naming the column names as described in the description provided. Lastly, we take care of missing values by assigning them NA.

Task 2

The variables that we will work with are the “Mean values of ten real-valued features computed for each cell nucleus”, “Tumor size”, and “Lymph node status” which are the predictors. The variable “Time” is our outcome variable.

Additionally, we need to code “Lymph node status” into a categorical variable with three levels: 0, 1-3, 4 or more. This is done by using the cut function which divides the continuous variable into factors/categorical variables by the given levels specified in the task description.

Lastly, we need to create a subset of the original dataset with 198 observations to only contain columns 3-13, columns 34 and 35. These columns correspond to the previously stated variables of interest. Note further that the subdataset only contains rows where the column “Outcome” have the value “R” which indicates recurrence. After doing this, we are left with 47 observations where only one contains a “NA”.

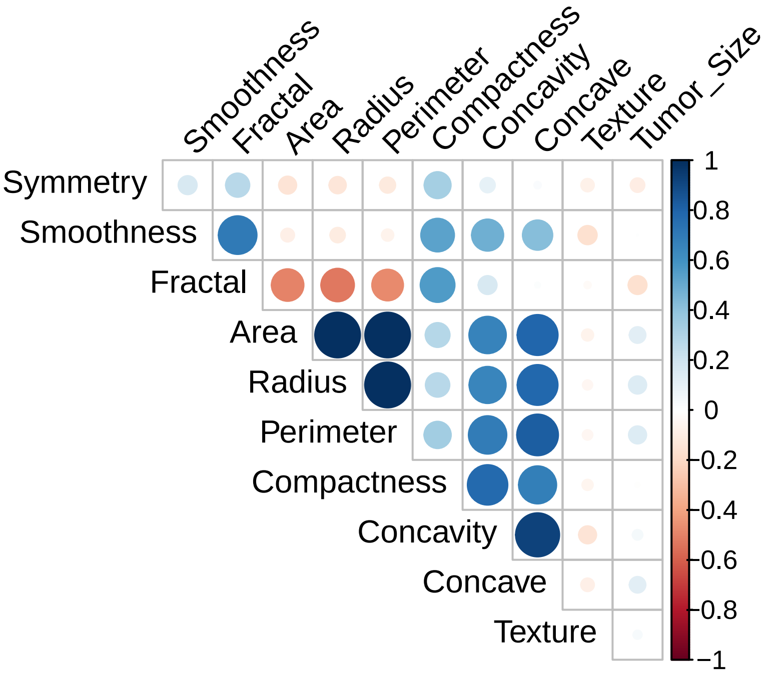
We proceed by giving some appropriate descriptive statistics and tables. First we give a table 1 containing summary statistics for each predictor are provided. Note that Lymph_Size contains one one missing value hence the mean is not reported.

Table 1: Descriptive Statistics

Feature	Min	X1st.Qu.	Median	Mean	X3rd.Qu.	Max
Time	1	9	16	25.09	36.5	78
Radius	12.34	15.81	19	18.4	20.29	27.22
Texture	14.34	19.23	21.49	21.78	24.05	30.99
Perimeter	81.15	104.9	123.7	121.6	133.7	182.1
Area	477.4	801	1104	1089.6	1289.5	2250
Smoothness	0.08217	0.09415	0.1034	0.10315	0.11175	0.1215
Compactness	0.06722	0.11345	0.1339	0.14272	0.16655	0.2363
Concavity	0.05253	0.11155	0.1655	0.16317	0.212	0.3368
Concave	0.03334	0.06807	0.08994	0.09394	0.10955	0.1913
Symmetry	0.1424	0.1722	0.1867	0.1879	0.1983	0.2356
Fractal	0.05025	0.05638	0.06082	0.06125	0.06508	0.07451
Tumor_Size	0.4	2.4	3	3.462	4	10
Lymph_Size	0:12	1:12	2:22	NA’s:1		

From the correlation plot, we observe there exists:

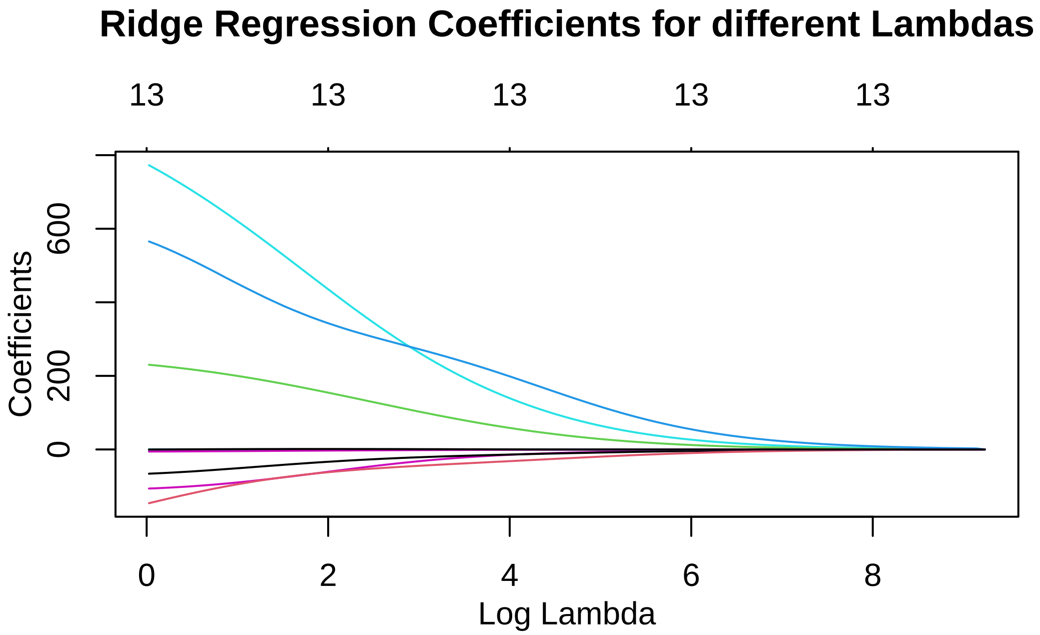
- a strong positive correlation for Area, Radius, and Perimeter. This makes intuitive sense since, if the radius of a tumor increases, so should the area and perimeter.
- a positive correlation between concavity and compactness. This means tumors that have more compactness also tend to have more concave features.
- a slight negative correlation between Fractal and Area (correspondingly also Radius and Perimeter). This suggests that as Fractal increases, the others decrease slightly.



The prescense of a strong correlations between variables such as Area, Radius, and Perimeter indicate the presence of multicollinearity, which can lead to inflated variance and unstable coefficient estimates. To address this issue, one approach is to apply regularization techniques such as Ridge Regression or Lasso Regression, which help shrink coefficients and mitigate multicollinearity, leading to more stable and reliable predictions.

Task 3

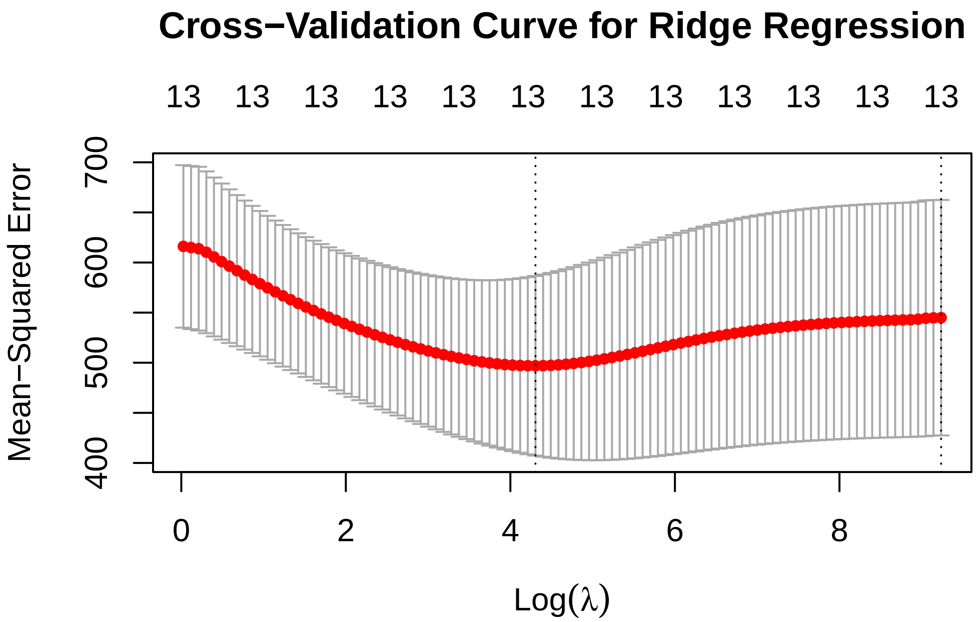
We will now train a ridge regression model to predict time to occurence (outcome variable) using the 12 selected features as predictors. For the λ parameter the default grid of values in the glmnet R function is used. We omitted one row where Lymph_size had a missing value in order to fit the model. The plot below illustrates how the coefficients of the predictors vary across different levels of the regularization parameter $\log(\lambda)$.



We have $\log \lambda$ on the x-axis and the coefficient values on the y-axis. Some of the coefficients start with values greater than 200 at $\log \lambda = 0$ and decreases towards 0 as λ increases. When λ is small, the model has minimal regularization, meaning the coefficients remain large. When λ increases, coefficients gradually decrease, eventually approaching zero. In this case, Ridge regression makes the coefficients shrink but never become exactly zero, meaning all variables still contribute but with reduced effect.

Task 4

We now perform a 5-five cross-validation to get the optimal λ value, which will help us minimize the MSE. The plot shows the MSE against the values of $\log(\lambda)$.



The red dots in the plot is the cross-validated MSE for each value of λ and the gray error bars gives the variability across the different folds. The left vertical dash line correspond to λ_{min}, the value that minimizes the MSE. The right dashed line correspond to λ_{1se}, which is the largest λ within one standard error of the minimum. Note that $\lambda_{\text{min}} = 74.05392$ for which we obtain the best predictive performance. The table below reports the coefficients of the predictors for the optimal lambda (λ_{min}) value.

Table 2: Coefficients at optimal lambda

Coefficient	Value at lambda.min
(Intercept)	21.6473159
Radius	-0.3942953
Texture	-0.1351632
Perimeter	-0.0575222
Area	-0.0028878
Smoothness	111.6713884
Compactness	-11.4032987
Concavity	-12.1123326

Coefficient	Value at lambda.min
Concave	-28.0338287
Symmetry	47.6410978
Fractal	173.0877826
Tumor_Size	-0.1590282
Lymph_Size1	-0.6911669
Lymph_Size2	0.0915845

The final coefficients at the optimal lambda (λ_{min}) from the ridge regression model illustrate the contribution of each predictor to the outcome while addressing multicolinearity. Note that most coefficients are small in magnitude, indicating that ridge regression effectively reduces them to mitigate overfitting. However, variables such as Smoothness (111.67), Fractal (173.09), and Symmetry (47.64) retain larger absolute values, indicating they have a stronger association with the response variable.

Task 5

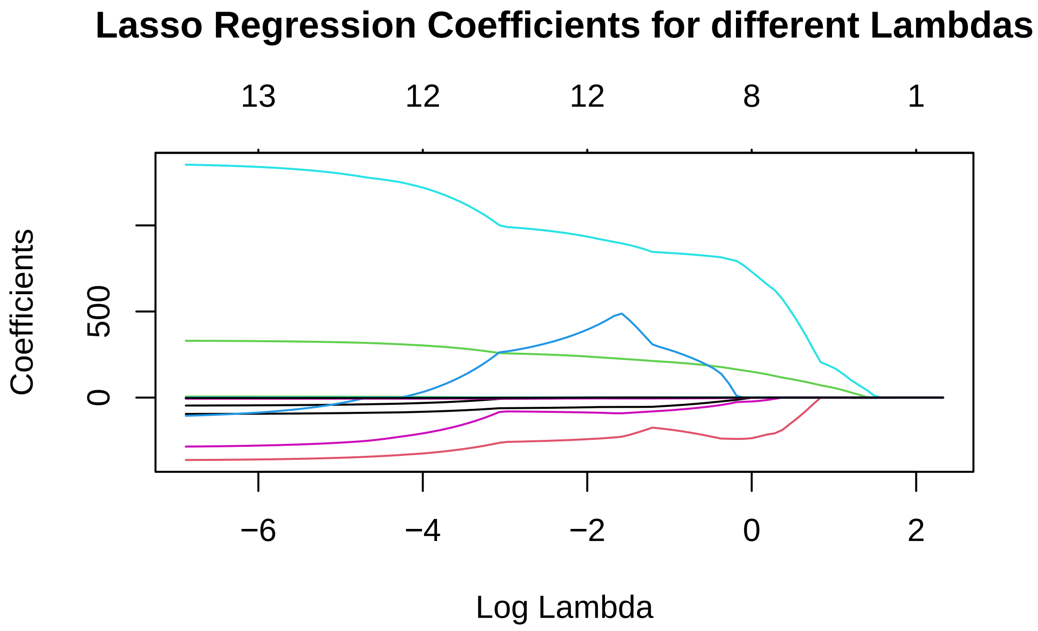
We calculate and report the MSE on the whole set of the recurrent group for the model using the optimal lambda value chosen previously. The formula for MSE

$$\frac{1}{n}\sum_{i}^{n}\left(y_i - \hat{f}(x_i)\right)^2,$$

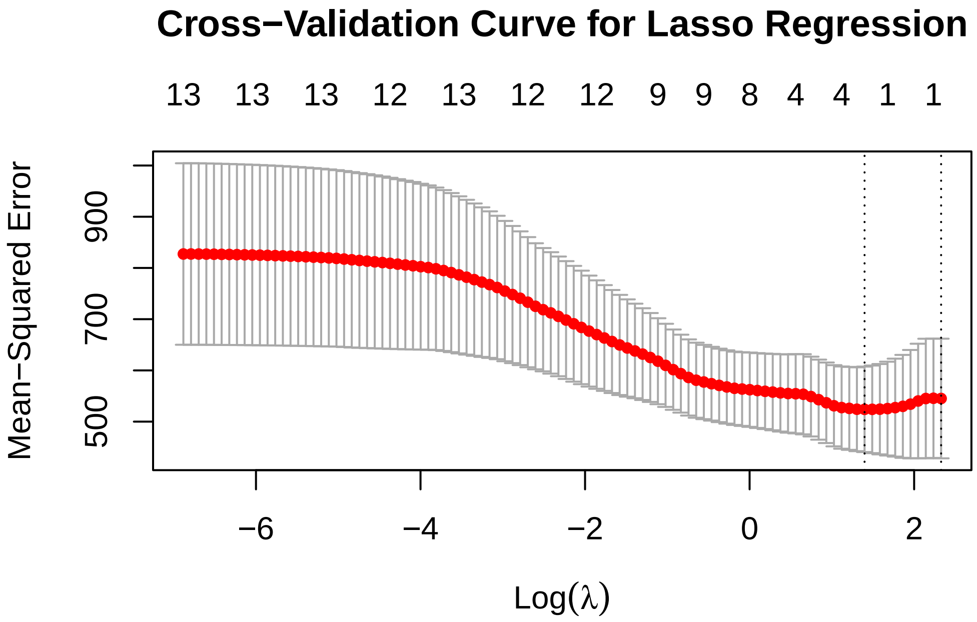
where y_i is the observed recurrence time and $\hat{f}(x_i)$ is the predicted one. The computed MSE is 400.2527.

Task 6

We now redo tasks 3-5 using the Lasso regression model. Similarly, Lasso regression model is trained for the prediction of time to recurrence (outcome variable) using the other 12 features as predictors. For the λ parameter the default grid of values in the glmnet R function is used. Below the plot depicts the coefficients of the predictors for different levels of the regularization parameter $\log(\lambda)$.



The x-axis shows $\log \lambda$ and the y-axis again represents the coefficient values. Note that as λ increases, coefficients shrinks to 0, gradually leading to a model with less predictors. Now we will perform a cross validation to identify the optimal λ and the resulting features that were not set to zero.



The cross-validation plot from cv.glmnet for Lasso Regression illustrates the relationship between the logarithm of the regularization parameter ($\log(\lambda)$) and the mean squared error (MSE). Similar to the previous plot with Ridge, the red dots represent the cross-validated MSE for each value of λ, while the gray error bars indicate the variability across different folds. The left vertical dashed line corresponds to λ_{min}, the value that minimizes MSE. We choose λ_{min} as the resulting coefficients provide the best predictive performance.

Table 3: Non-Zero Coefficients at Optimal Lambda for Lasso Regression

Coefficient	Value at lambda.min
Intercept	54.189931
Radius	-1.847190
Smoothness	44.457887
Symmetry	3.481176

Note that many parameters shrank to 0 at the optimal λ_{min}. Notably, the variables Texture, Perimeter, Area, Compactness, Concavity, Concave, Fractal, Tumor Size, Lymph Size (level 1), and Lymph Size (level 2) are excluded from the model. Also Lymph Size (level 0) is absorbed into the intercept as part of the encoding.

We now calculate and report the MSE on the whole set of the recurrent group for the model using the optimal lambda value. Using the same formula as before in task 5. The computed MSE is 412.5034.

Task 7

We achieve a smaller MSE value of 400.2527 using Ridge Regression compared to 412.5034 for Lasso Regression. Basing on these values, it seems that Ridge performed slightly better in terms of predictive accuracy on this dataset. However, the difference is not substantial and both models have relatively high MSE values, indicating that there is much room for improvement.

A more rigorous approach to comparing the performance of the two prediction methods is nested cross-validation. This technique consists of an outer loop of cross-validation to assess the generalization performance of each method and an inner loop within each fold of the outer loop to optimize the λ parameter.

Task 8

Some considerations to account for before using the trained model(s) for predicting the time to recurrence using the values of the predictors are:

- Selection bias: The dataset we have is limited to patients who have already undergone surgery. So the model may not generalize well to patients with different treatment histories or tumor characteristics.
- Small dataset: The number of patients we have in the data is small, which could have limited the ability to train a robust model.
- Including additional features (variables not included in Task 2). Including these features could potentially provide more information about the variability of the cell nuclei, which could improve predictive performance.