Applying LASSO Regression for Multinomial Cancer Classification*

Harrison Jones 1006697335 Liam Wall 1007674991

December 8, 2024

In this paper our team, `Whaddyat?`, competed in the Kaggle cancer type classification competition. After developing a model, we submitted a prediction that earned us third place with a prediction score of 1.000.

1 Problem Statement

1.1 Research Question

The objective of this study is to develop a model that can accurately classify cancer patients into three distinct subtypes using a gene expression dataset containing over 12,000 gene expression variables for 886 cancer patients.

1.2 Introduction

Cancer type classification based on gene expression is crucial for understanding the genetic differences between cancer types. Accurate classification models can aid patient outcomes by improving clinical decision making, helping to guide personalized treatment plans, and further our knowledge of cancer biology. By utilizing gene expression data, we aim to develop a model that can accurately predict whether a patient was diagnosed with one of three types of cancer: Glioblastoma Multiforme (GBM), Lung Squamous Cell Carcinoma (LUSC), and Ovarian Cancer (OV).

This research plays an important role in the medical field because cancer remains a leading cause of death worldwide, with an estimated 20 million new cases and 9.7 million deaths in 2022 alone. Approximately one in five people develop cancer in their lifetime, making early

*Code and data are available at: https://github.com/Lwall02/Cancer-Type-Classification.

detection and effective treatment essential. (Organization 2024) For cancers like GBM, LUSC, and OV, this research holds particular promise. These cancers not only pose substantial challenges due to their aggressive nature but also stand to benefit greatly from genetic testing and targeted therapies. Understanding the genetic distinctions between these cancer types will lead to innovative medical breakthroughs, and the first step is developing an accurate predictive model.

1.3 Glioblastoma Multiforme (GBM)

GBM is a fast-growing and aggressive brain tumor with a poor prognosis. It has a survival rate in the first year post diagnosis of 40% and only about 17% after the second year. In adults, the five year survival rate is 5.6%. Typically, GBM requires surgery before any other treatment, making molecular/genetic testing common among those diagnosed. This testing is to identify specific genetic markers and help confirm tumor diagnosis, inform treatment options, and predict the patient's outcome. (Nuerological Surgeons 2024) (Association 2024) Developing an accurate predictive model for GBM could aid in early and accurate classification and lead to future innovations that ultimately improve survival outcomes.

1.4 Lung Squamous Cell Carcinoma (LUSC)

Lung cancers are the most common cause of cancer related deaths worldwide and LUSC makes up about one third of lung cancer related diagnoses. LUSC occurs when a tumor grows in the lung. These cancer cells can spread to other parts of the body, or metastasize, easily. For this reason, early detection increases the patient's chance of survival greatly. LUSC can not be diagnosed until the tumor cells are looked at under a microscope, which makes genetic testing essential among those diagnosed. This testing can reveal genetic abnormalities of the tumor and makes targeted therapy a promising option. 30% of lung cancers can be more effectively treated with these targeted therapies. (Publishing 2023) (Center 2021) Developing accurate predictive models for LUSC could help speed up the diagnosis and implementation of targeted therapies, improving patient outcomes.

1.5 Ovarian Cancer (OV)

Ovarian cancer develops when the cells of the ovary grow uncontrolled and form tumors. Nobody knows exactly what causes ovarian cancer. Approximately 15% of ovarian cancers are linked to genetic mutations. Clinical trials have shown that a specific targeted therapy, PARP Inhibitors, delay the progression of the cancer. Patients with certain genetic mutations are particularly susceptible to this. This encourages the research focusing on genetic therapies which are more precise and potentially more effective than other treatments. They also reduce the risk to normal tissues and have less side effects compared to chemotherapy. (Health 2024) (Medicine 2024) Developing an accurate predictive model for ovarian cancer, similar to GBM

and LUSC, could improve early detection and identify patients most likely to benefit from targeted therapies.

2 Statistical Analyses

2.1 Data

The data used in this study originates from a Kaggle dataset derived from The Cancer Genome Atlas (TCGA), a collaborative effort by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to map the key genomic changes in 33 types of cancer in over 11,000 patients. The TCGA dataset is publicly available and has contributed to more than a thousand studies of cancer by independent researchers and TCGA research network publications. (Kaggle 2024) As discussed, in this paper, we focus on three cancer types: the glioblastoma multiforme (GBM), the lung squamous cell carcinoma (LUSC) and ovarian cancer (OV).

Both the training and testing data contain 12,043 features of 886 patients in the training data and 379 patients in the testing data. To be specific, in the training data, 376 patients have GBM, 90 patients have LUSC, and 420 patients have OV. The features are all gene expression data collected using the Affymetrix HT Human Genome U133a microarray platform, which was processed by the Broad Institute of MIT and Harvard University's cancer genomic characterization center. (Kaggle 2024) Gene expression measures how actively a gene is being transcribed into RNA, providing insights into the functional activity of genes in different tissues. In this dataset, these values are log-transformed for better interpretability.

In this paper, all models were developed using only the training data because the available testing data excludes the column identifying which type of cancer each patient has. More on the cross validation techniques used is discussed later. All data cleaning and analysis was done using the open source statistical programming language R (R Core Team 2023). The packages `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `testthat` (Wickham 2011), `arrow` (Richardson et al. 2023), `glmnet` (Friedman, Tibshirani, and Hastie 2010), and `coefplot` (Lander 2022), were also used for data cleaning, testing, analysis, and model development.

2.1.1 Data Cleaning

The data cleaning process focused on ensuring that column names were properly formatted and compatible with autofill functionality. We also verified that the dataset was free of missing values and inconsistencies. Additionally, the cancer column in the training data was converted into a three-level factor with values 1, 2, and 3 to allow for more effective handling by the model.

2.1.2 Data Transformation

We did not employ any data transformation ourselves on this data, as the downloaded training and test datasets already contained gene expression values in log space. This transformation is particularly advantageous for our use of the glmnet model, which applies Ridge and Lasso regression. Standardization of input variables is essential for effective regularization in these models. The log transformation minimizes the influence of outliers, stabilizes variance, and promotes a more normal distribution of the data. Figure 1 represents the distribution of log-transformed gene expression values across all genes in the dataset. While individual genes may exhibit normality in log space, this plot shows the combined contributions of diverse gene expression profiles, emphasizing the variability in gene activity among samples. Importantly, interpretability is preserved with log-transformed data, as a one-unit increase or decrease in log values corresponds to a multiplicative change on the original scale.

2.1.3 Feature Selection

No explicit feature selection techniques were applied as the glmnet model inherently performs regularization and variable selection during training. All 12,043 features were employed in the building of our model.

2.2 Model

2.2.1 Model Choice

We used a Multinomial Logistic Regression implemented through the `cv.glmnet` function, a method well-suited for high-dimensional datasets where the number of parameters features far exceeds the number of samples. This method is advantageous in this context of genetic data, where the vast majority of features may be irrelevant or redundant. The LASSO approach, which penalizes the absolute value of coefficients, is particularly effective for feature selection, as it encourages sparsity by shrinking the coefficients of irrelevant features to zero. This inherent feature selection capability makes LASSO an ideal candidate for our classification task. In addition, the flexibility of the `cv.glmnet` function allows us to compare LASSO and Ridge regularization approaches, allowing us to determine which method best fits the data and enhances model performance. (Tay, Narasimhan, and Hastie 2023)

2.2.2 Cross Validation

To optimize the regularization parameter, (lambda), and to determine the most appropriate regularization method, we employed 10-fold cross-validation. Cross-validation is a critical technique for assessing the model's ability to generalize to data, reducing the risk of overfitting. By performing cross-validation on a wide range of regularization parameters and comparing

the performance of both LASSO and Ridge, we ensure that the selected model achieves optimal performance without overfitting to the training set.

2.2.3 Model Justifiation

We chose `cv.glmnet` function due to its ability to handle datasets with a large number of predictors efficiently, providing a versatile framework for regularized regression. The ability to perform LASSO and Ridge regularization allows for flexibility in addressing the complex genetic data. The regularization of LASSO is particularly beneficial for sparse models, where many of the features may be irrelevant or redundant. On the other hand, regularization of Ridge works well when it is assumed that most features contribute to the model, allowing for shrinkage without eliminating any features. By utilizing the cv.glmnet function without explicitly specifying whether to use LASSO or Ridge, we ensured an unbiased comparison of both methods to determine the most appropriate model for our dataset.

2.2.4 Training and Evaluation

The model was trained using the given training data to predict the cancer subtype outcome (multinomial classification). Once the optimal value was identified via cross-validation, look at Figure 2, the final model was used to generate predictions on the testing data. Model performance was evaluated based on its ability to accurately classify patients into the correct cancer subtypes. The effectiveness of the model was gauged by measuring the classification accuracy, ensuring that the selected model generalized well to new data. This iterative approach of training and validation allowed for the identification of the model configuration that best-balanced bias and variance, ultimately leading to robust predictions. This process of regularization selection and cross-validation ensures that the model is both accurate and generalizable.

2.3 Evaluation Metrics

The primary metric used to evaluate the performance of our models is the testing misclassification error, as it directly reflects the model's ability to correctly classify unseen data. Misclassification error is calculated as the proportion of incorrect predictions to the total number of predictions, with a lower value indicating better model performance. This metric is particularly relevant in the context of our study, where the goal is to classify cancer patients into three distinct subtypes based on gene expression data. A model that minimizes the misclassification error will be more reliable and applicable in clinical settings, where accurate classification is critical for patient care.

Although misclassification error provides a comprehensive measure of performance, we are also interested in selecting a model that is both effective and interpretable. In scenarios where

multiple models achieve similar classification accuracy, simplicity and interpretability become important factors in model selection. A model that achieves similar predictive performance but with fewer variables or a clearer structure is preferable, particularly when considering real-world applications. Simplifying the model enhances its usability and interpretability, which is critical for those who need to understand the decisions made by the model.

Given this, while we will first focus on the misclassification error to determine which model offers the best classification accuracy, we will also consider the complexity and interpretability of the models. If multiple models achieve comparable results in terms of misclassification error, the less complex and more interpretable model will be favored. In particular, LASSO regularization, with its ability to perform feature selection and identify the most relevant genes, may offer valuable insights into the key factors contributing to the cancer classification. Therefore, while the misclassification error is the most important evaluation criterion, model complexity, and interpretability will be key considerations in our final selection process should similar misclassification errors arise in separate models.

3 Results and Conclusion

3.1 Model Performance

The model developed using multinomial logistic regression with regularization through cross-validation achieved 100% classification accuracy on the testing dataset. This perfect prediction performance suggests that the model is highly effective in distinguishing between the three cancer subtypes based on the gene expression profiles of the patients. The training and testing process involved rigorous cross-validation, which allowed for optimal tuning of the regularization parameter , ensuring the model's ability to generalize well to unseen data. The accuracy of the predictions on the testing dataset further supports the robustness of the model, as it indicates the model's consistent ability to correctly classify each cancer subtype without overfitting the training data.

Both LASSO and Ridge models were evaluated using the same dataset, and interestingly, both models achieved identical performance in terms of classification accuracy on the testing data, with 100% correct classification. This outcome displays the strength of both regularization techniques in handling high-dimensional data. These results suggest that both methods are highly effective in this context, managing to identify the underlying patterns and relationships in the gene expression data with remarkable precision.

3.2 Conclusion

The results from this analysis demonstrate that multinomial logistic regression, when coupled with regularization techniques such as LASSO and Ridge, can yield highly accurate models for

the classification of cancer subtypes based on gene expression data. The perfect classification rates observed with both LASSO and Ridge models provide evidence of the model's accuracy and reliability in distinguishing between cancer subtypes, even in the face of high-dimensional data.

Although both LASSO and Ridge models achieved the same level of performance in terms of classification accuracy, LASSO's inherent feature selection capabilities make it particularly valuable. LASSO's ability to shrink coefficients to zero for less relevant features allows for a more interpretable model, enabling the identification of the most influential genes involved in cancer subtype classification. This interpretability is an advantage, as it provides insights into the genetic markers that contribute to the differentiation of cancer subtypes. This model could facilitate further research into the molecular underpinnings of cancer and inform the development of targeted therapies.

Given the excellent performance of the models and the interpretability offered by LASSO, this approach holds significant potential for deployment in clinical settings. With continued validation and refinement, such a model could support more accurate cancer diagnoses, inform personalized treatment plans, and contribute to a deeper understanding of the genetic factors driving cancer progression.

4 Discussion

4.1 Limitations

While the model achieved perfect classification accuracy on the test data, it is essential to acknowledge several limitations that may impact the generalizability of these findings. One of the primary concerns is the potential for overfitting, especially given the high-dimensional nature of the dataset. In this study, the number of features greatly exceeds the number of samples, which increases the risk of the model fitting to noise or spurious patterns present in the data. Although cross-validation was employed to mitigate this risk, the simplicity of the dataset and relatively small sample size in comparison to the number of features still raise concerns about the model's generalizability to other populations.

Another important limitation lies in the lack of external validation. While the model exhibited perfect performance on the provided test dataset, its ability to generalize to data from different sources remains unknown. External validation using independent datasets—preferably with diverse cancer populations—would be necessary to assess the true robustness of the model and its ability to maintain high classification accuracy across varying conditions. Without this validation, there is a risk that the model's performance may not be representative of its practical applicability in diverse clinical settings.

Furthermore, the gene expression profiles used in this study may not fully capture the complex nature of cancer. Cancer subtypes are influenced by a wide range of genetic, clinical, and

environmental factors. The dataset used in this analysis is limited to gene expression data and does not include other potentially significant information such as demographics, clinical history, or environmental exposures. This limitation may reduce the model's ability to account for all potential variations in cancer subtypes, as these non-genetic factors can also play a significant role in tumor progression and differentiation.

4.2 Future Directions

Given the results of this, several avenues for future research and model improvement are worth exploring. One potential direction is the investigation of more complex machine learning models, such as deep learning architectures or ensemble methods (e.g., random forests, gradient boosting), which may improve classification accuracy and enhance the model's ability to generalize to more complex datasets. These methods, while computationally more intensive, have the potential to capture nonlinear relationships and higher-order interactions within the data that simpler models like logistic regression may miss.

Another promising direction is to integrate additional forms of clinical data into the model. By incorporating features such as patient demographics, clinical history, and treatment responses, the model could achieve a more holistic view of cancer progression. This multimodal approach could potentially improve classification accuracy and make the model more robust, as it would consider a broader range of factors beyond gene expression. Moreover, combining genomic and clinical data has the potential to yield insights into the complex interplay between genetics, environment, and disease, providing a more nuanced understanding of cancer subtypes.

Finally, external validation of the model is a critical step in assessing its practical applicability in real-world scenarios. Performance evaluation on independent and diverse datasets from different geographic regions, cancer cohorts, and medical institutions is essential to ensure the model's robustness. Such external validation will help identify potential biases in the data, assess the model's ability to generalize across populations, and confirm whether the model's 100% classification accuracy on the testing data can be replicated in broader, more varied contexts.

In summary, while the model developed shows promise, addressing the limitations discussed above and exploring these future research directions will be essential for improving its robustness, interpretability, and practical application. By extending this work to incorporate additional data sources and more complex modeling techniques, future research can contribute significantly to the development of more accurate and reliable cancer classification tools, ultimately enhancing patient outcomes and advancing personalized cancer treatment strategies.

5 Figures

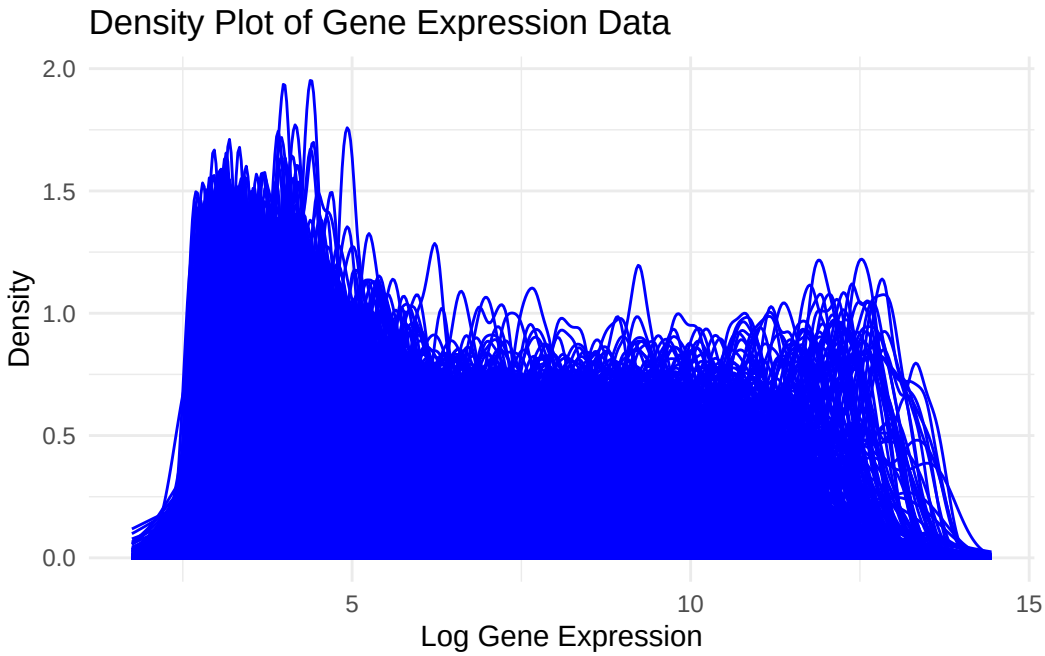

Figure 1: This plot represents the distribution of log-transformed gene expression values across all genes in the dataset. While individual genes may exhibit normality in log space, the aggregated density plot suggests multimodal tendencies and a slight right skew.

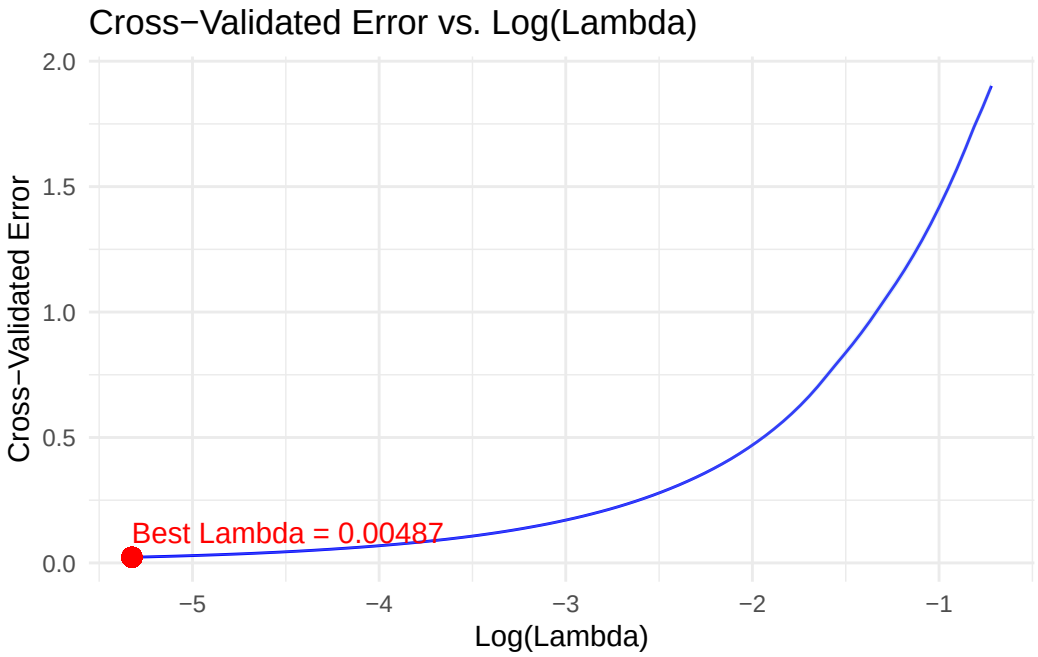

Figure 2: This plot represents the cross-validation error of our `glmnet` multinomial model as our lambda regularization parameter increases. We also show the chosen lambda at =0.00487. This lambda value is chosen such that the model acheives the best fit, balancing its bias and variance.

6 Code

All the code needed to reproduce this paper and the results of our model can be found in our Github repository https://github.com/Lwall02/Cancer-Type-Classification. It can also be found below knitted to this paper.

Specifically, the code that allows you to download, clean, test and produce the model is available in the `scripts` folder. The model is saved as a .rds file in the `models` folder. The raw and cleaned training and testing data as well as our prediction results is available in the `data` folder. Lastly, the code and references used to make this paper is all available in the `paper` folder. The `testingresult.csv` file in the `data` folder is our exact submission for the Kaggle competition that earned us third place with a score of 1.000.

```
#### Workspace setup ####
library(tidyverse)
library(janitor)
library(arrow)

#### Read in data ####
training_data <- read_csv("data/raw_data/train.csv")
test_data <- read_csv("data/raw_data/test.csv")

#### Clean data ####
# Clean column names
clean_training_data <- training_data |>
  clean_names()
clean_test_data <- test_data |>
  clean_names()

# Convert the response variable to a factor
clean_training_data$cancer <- factor(clean_training_data$cancer, levels = c(1, 2, 3))

#### Save data ####
write_parquet(clean_training_data, "data/analysis_data/clean_training_data.parquet")
write_parquet(clean_test_data, "data/analysis_data/clean_test_data.parquet")
```

```
#### Read in clean data ####
# Step 1: download data and get features
clean_training_data <- read_parquet("data/analysis_data/clean_training_data.parquet")
clean_test_data <- read_parquet("data/analysis_data/clean_test_data.parquet")

# If needed: load the raw training and testing data
```

```r
# training_data <- read.csv("train.csv")
# testing_data <- read.csv("testing.csv")


#### GLMnet model and predictions####
set.seed(9)

train_features <- clean_training_data[, -c(1:2)]  # Remove first two columns from training
test_features <- clean_testing_data[, -c(1:1)]    # Remove first column from testing data

# Perform cross-validation to find optimal lambda
# Use training data for cross-validation, predicting the 'cancer' outcome variable
cvfit <- cv.glmnet(as.matrix(train_features), clean_training_data$cancer, family = "multin

# To check the coefficients at the best lambda
cat("Coefficients at the best lambda:\n")
print(cvfit$nzero[which(cvfit$lambda == cvfit$lambda.min)])

coefs_list <- predict(cvfit, type = "coef", s = "lambda.min")

# Loop through each class
for (i in 1:3) {
  coefs_sparse_matrix <- coefs_list[[i]]

  coefs_dense_matrix <- as.matrix(coefs_sparse_matrix)

  cat("Number of coefficients for Class", i, ":", nrow(coefs_dense_matrix), "\n")

  non_zero_class_coefs <- coefs_dense_matrix[coefs_dense_matrix[, 1] != 0, ]

  cat("\nClass", i, "Coefficients:\n")
  print(non_zero_class_coefs)
}

# Ensure the columns of testing data match the training data
# Reorder testing data columns to match the training data's feature order
test_features <- test_features[, colnames(train_features)]

# Predict the class labels using the best lambda for the testing data
predictions_test <- predict(cvfit, newx = as.matrix(test_features), s = "lambda.min", type
```

```r
# Check the predictions
head(predictions_test)

# Create the new matrix with ID and predicted cancer labels
predicted_result <- data.frame(ID = clean_test_data$id, Predicted_Cancer = predictions_tes

# Display the result
cat("Predicted Results (ID and Predicted Cancer):\n")
print(predicted_result)


#### Save prediction results and model ####
write.csv(predicted_result, "data/prediction_data/testingresult.csv", row.names = FALSE)
saveRDS(
  cvfit,
  file = "models/glmnet_model.rds"
)
```

References

Association, American Brain Tumor. 2024. *Glioblastoma (GBM)*. https://www.abta.org/tumor_types/glioblastoma-gbm/.

Center, MD Anderson Cancer. 2021. *Do Lung Cancer Patients Need Molecular Profiling?* https://www.mdanderson.org/cancerwise/do-lung-cancer-patients-need-molecular-profiling-before-treatment.h00-159457689.html.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Friedman, Jerome, Robert Tibshirani, and Trevor Hastie. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

Health, National Institute of. 2024. *Advances in Ovarian Cancer Research*. https://www.cancer.gov/types/ovarian/research.

Kaggle. 2024. *Classification of Cancer Type*. https://www.kaggle.com/competitions/classification-of-cancer-type/overview.

Lander, Jared P. 2022. *Coefplot: Plots Coefficients from Fitted Models*. https://CRAN.R-project.org/package=coefplot.

Medicine, John Hopkins. 2024. *Ovarian Cancer*. https://www.hopkinsmedicine.org/health/conditions-and-diseases/ovarian-cancer.

Nuerological Surgeons, American Association of. 2024. *Glioblastoma Multiforme*. https://www.aans.org/patients/conditions-treatments/glioblastoma-multiforme/.

Organization, World Health. 2024. *Global Cancer Burden Growing, Amidst Mounting Need for Services*. Lyon, France; Geneva, Switzerland. https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services#:~:text=In%202022%2C%20there%20were%20an,women%20die%20from%20the%20disease.

Publishing, Harvard Health. 2023. *Squamous Cell Carcinoma of the Lung*. https://www.health.harvard.edu/cancer/squamous-cell-carcinoma-of-the-lung-a-to-z.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2023. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Tay, J. Kenneth, Balasubramanian Narasimhan, and Trevor Hastie. 2023. "Elastic Net Regularization Paths for All Generalized Linear Models." *Journal of Statistical Software* 106 (1): 1–31. https://doi.org/10.18637/jss.v106.i01.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.