Common Cancer Surgeries: A Statistical Analysis of Surgical Trends in California Hospitals

MATH 4322

Group 14: Marwan Aridi & Hasher Khan

April 29, 2024

I. Introduction

- **Description of Data:**
    - Using data from the Department of Health Care Access and Information from the State of California, which includes records of cancer procedures in general acute care hospitals from 2013 to 2022, this research examines the effects of several factors on the number of cancer cases. The dataset offers a thorough understanding of cancer treatment patterns over a ten-year period by containing precise information on each county in the state of California, every hospital in the state, the type of surgeries, and the number of cases for each type of surgery.
- **Description of Variables:**
    - Column 1: *Year*
    - Column 2: *County*
    - Column 3: *hospital*
    - Column 4: *Surgery* (Types of cancer surgeries like breast, brain, lung, etc.)
    - Column 5: *# of Cases*
    - Column 6: *LONGITUDE*
    - Column 7: *LATITUDE*
- **Data Question and Inspiration:**
    - **What factors/variables affect the number of cases the most?** If the year is found to have the greatest influence, this implies that variations in the number of cancer cases across time may have been mostly caused by annual variations. This might be because, in certain years relative to others, a greater number of persons

were receiving diagnoses; which could be because of things like improved screening initiatives, raised awareness, or population growth. Conversely, if surgery type has a larger impact, this means that different surgeries are linked to varying numbers of cancer occurrences, which may be a reflection of the prevalence of particular cancer kinds. We will utilize two techniques to get answers: Linear Regression and Random Forest. We may observe clear correlations between the number of cancer cases and other variables, such as Years or the type of surgery, with the use of linear regression. It is simple to understand and will provide us with a clear picture of the situation. Conversely, Random Forest is a little more complex. It examines every possible way that variables could interact to impact cancer cases, preventing us from overlooking anything crucial because it's advanced.

II. Methods and Results

- **Linear Regression Model:**
    - Because linear regression is particularly effective at highlighting the relationship between a continuous result variable and several predictor factors, we decided to use it in our research. In this study, we explore how various factors impact the continuous variable of the number of cancer cases. We can measure the precise amount that every factor changes the number of instances and determine which factors have a statistically significant influence using linear regression. Linear regression is an effective method for addressing our primary concern because of its simplicity and clarity in analyzing the results.
    - Linear Regression model has its advantages and disadvantages. The advantage of using Linear Regression is the simplicity. For instance, Linear Regression is

very straightforward to work with using R and only requires a few lines to get an output of our result, making it less complex. Another advantage of using Linear Regression is interpretability. For example, the coefficients in the summary function in R provide a clear correlation. That is, it shows our unit might increase or decrease the number of cancer cases. The disadvantage of using the Linear Regression model is outliers. Outliers can actually affect the slope of the slope line of the model and it may give us a poor fit for the data. Another disadvantage of using Linear Regression is linearity. Linear Regression is based on the assumption that the relationship between independent variables and the dependent variable has to be linear. When the true relationship is not linear, the model cannot possibly capture all features of the data, which would result in incorrect predictions or understanding.

- Now, we will start with the linear regression model of the database (comparing all four predictors): Our model formula would be:

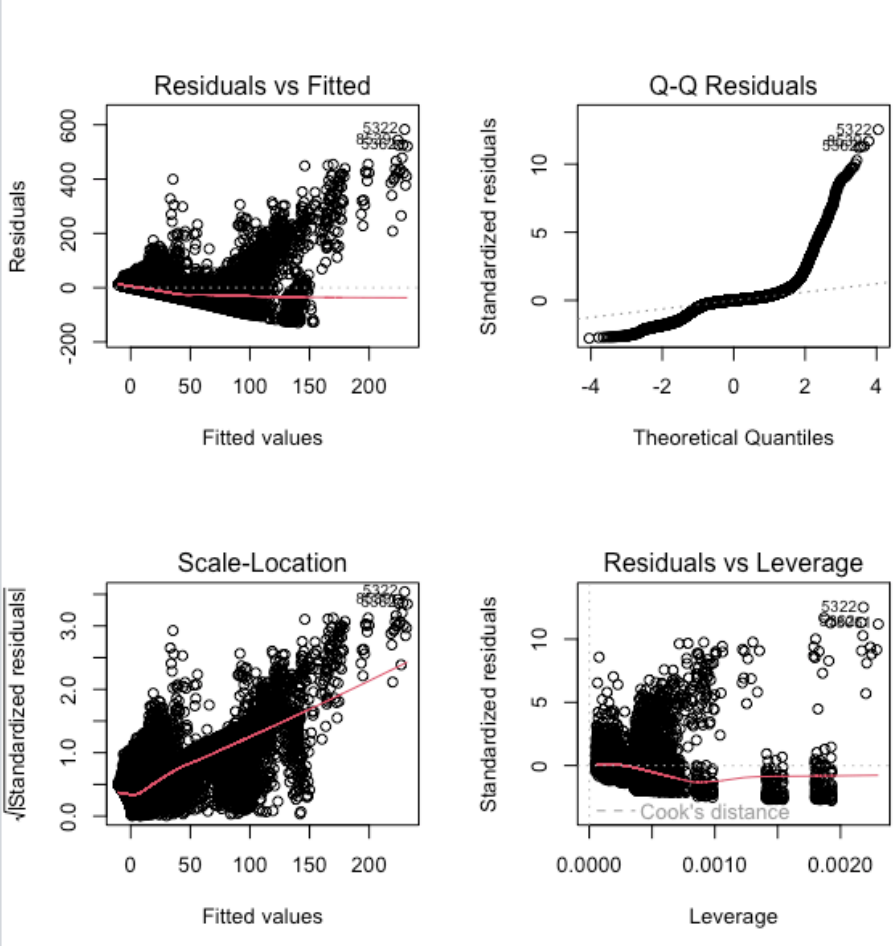
$$Y = \beta_0 + \beta_1 Year + \beta_2 Hospital\ Effect + \beta_3 Surgery\ Effect + \beta_4 County\ Effect + \epsilon$$

```
Call:
lm(formula = `# of Cases` ~ Year + hospital_effect + surgery_effect +
    county_effect, data = data_model)

Residuals:
    Min      1Q  Median      3Q     Max
-129.45  -10.29    0.39    9.08  583.23

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.476e+03  2.370e+02  -6.225 4.90e-10 ***
Year             7.276e-01  1.175e-01   6.193 6.01e-10 ***
hospital_effect  4.109e-03  6.452e-05  63.688  < 2e-16 ***
surgery_effect   6.095e-03  6.569e-05  92.780  < 2e-16 ***
county_effect    1.784e-03  1.724e-04  10.348  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46.59 on 19356 degrees of freedom
Multiple R-squared:  0.3998,    Adjusted R-squared:  0.3997
F-statistic:  3224 on 4 and 19356 DF,  p-value: < 2.2e-16
```

- Looking at the summary model above, we see that the F-statistic p-value is less than 0.05, so we can confidently conclude that we have a strong relationship, and reject the null hypothesis. In addition, all of the coefficients in the model are also statistically significant and reject the null hypothesis. Now, moving to the R-Squared and Adjusted R-Squared values, both sit at about 0.40, meaning that about 40% of the variation found in the dependent variable (# of Cases) can be explained by the model. This means that about 60% of this variability is due to either factors outside of the dataset, or due to variables that may not be inherent. While this does not necessarily suggest a weak fit for the data; this is most certainly not a very strong correlation either.
- Finally, we can use the summary for the linear regression model, in order to answer which variable will have the largest effect on the response variable, the "# of Cases." The coefficient for the variable, *Year*, is 0.7276. This indicates that this variable has a very significant and continuous increase in the number of cases as each year passes. Now, *surgery_effect* has a coefficient of 0.006095, which although is smaller, represents the largest immediate impact among the hospital, surgery, and county effects on the number of cases in a year. This suggests that although *hospital_effect* and *county_effect* also contribute to the changes in case numbers, the different cancer surgery types are directly correlated with the fluctuations in the case numbers. This means that according to this linear regression model, *Year* has the largest long-term influence on the number of cases because of its accumulating impact, and *surgery_effect* has the most impactful immediate effect on the number of cases annually.

- The predictors all demonstrate a statistically significant relationship to the $\#\ of\ Cases$. Therefore, its formula would be:

$$\#\ of Cases\ =\ -1.476\ \times 10^{3} + 0.7276\ \times Year + 4.109\ \times 10^{-3} Hospital\ Effect\ + 6.095\ \times 10^{-3} Surgery\ Effect\ +\ 1.784\ \times 10^{-3} County\ Effect$$



- The residual plots of the linear regression model pose some major concerns. The Residuals vs Fitted and Scale-Location plots both indicate heteroscedasticity, which is shown by the increase in the variance of residuals with increasing fitted values, and this means that the equal variances assumption does not hold. The Normal Q-Q plot shows that the residuals are not normally distributed and have slight deviations from normality towards the right tail of the plot. And for the final plot, the Residuals vs Leverage, there are a multitude of outliers and high

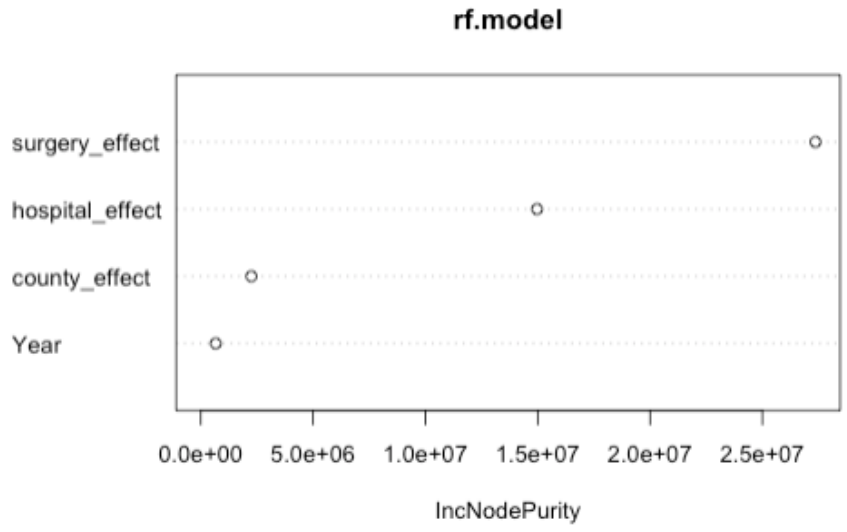
leverage points. These residual plots are all indicative of the fact that a linear regression model may not be adequate in looking at the variability of the response variable, *# of Cases.*

- **Random Forest Model:**
    - We picked a random forest model because of how well-suited it is when it comes to its ability to find and interpret non-linear relationships and interactions between variables. Most importantly, a random forest model offers a direct assessment of the importance of a variable, giving a clear look at which variables or factors are the most significant when it comes to determining the number of cases; which is the key question we are trying to answer. On top of this, the model has a higher resistance to overfitting because it averages the results from multiple trees, and it is also indifferent to heteroscedasticity, which, as we saw with the linear regression model, is a big plus. Despite the apparent trade-off when it comes to interpreting this model compared to a simpler model such as linear regression, a Random Forest can deliver a solid comprehensive understanding of the predictors' impacts, making it a great option for answering the question. In addition, Random Forest is very “flexible” since it works with both classification and regression problems. However, Random Forest might not be the best to work with sometimes because the more data size we have, the less it will perform. Lastly, Random Forest may require more memory and they tend to be slow especially when training and predicting the models.

```
# Fit the Random Forest model using the aggregated variables
rf.model <- randomForest(`# of Cases` ~ Year + hospital_effect + surgery_effect +
county_effect, data = data_model, ntree = 500)

# Print model summary
print(rf.model)

# Plot variable importance
importance <- randomForest::importance(rf.model)
print(importance)
randomForest::varImpPlot(rf.model)
```

```
Call:
 randomForest(formula = `# of Cases` ~ Year + hospital_effect +      surgery_effect +
county_effect, data = data_model, ntree = 500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 799.3293
                    % Var explained: 77.89
                IncNodePurity
Year                 665573.9
hospital_effect    14970265.7
surgery_effect     27365813.1
county_effect       2251357.6
```

- Looking at the summary for the random forest model above, one of the first results that we can see is the mean squared residual, which is 799.3293. This value shows that the model will deviate by about 799 cases compared to the model's prediction. The percentage of variance is sitting at 77.89%, which means that 77.89% of the variability in the data can be explained by the variables. Also, the variance being over 50% means that this model is a very good fit as it is able to explain most of the variance. The next values that we see are the importance values, which can also be seen in the importance plot below:



- Here, we can see that *surgery_effect* has the highest importance value, then *hospital_effect*, *county_effect,* and finally *Year.* This means that when it comes to the variability in the response variable, *# of Cases*, the different types of cancer surgeries that occur have the greatest impact on the number of cancer cases. When it comes to the last variable, *Year,* has the lowest importance here, and this is most likely due to the fact that when comparing the relationships and interactions of this variable to the other variables in the model, *Year* was most likely overshadowed by the other predictors. This indicates that while *Year* does have an effect on the variability in the number of cancer cases, its effect is much smaller than that of *surgery_effect* and *hospital_effect.*

III. MSE Comparison of Linear Regression & Random Forest Model

```
set.seed(10)

# Set up 10-fold cross-validation
ctrl <- trainControl(method = "cv", number = 10)

# Fit the linear model using caret for cross-validation
linear_model_cv <- train(
  `# of Cases` ~ Year + hospital_effect + surgery_effect + county_effect,
  data = data_model,
  method = "lm",
  trControl = ctrl
)



#-----------------------------------------------------#
```

```r
library(ranger)

# Set up k-fold cross-validation manually
k <- 10
folds <- cut(seq(1, nrow(data_model)), breaks = k, labels = FALSE)

# Perform 10-fold cross-validation
results <- data.frame(MSE = rep(NA, k))


for(i in 1:k){
  # Segment data into training and testing
  testIndexes <- which(folds == i, arr.ind = TRUE)
  testData <- data_model[testIndexes, ]
  trainData <- data_model[-testIndexes, ]

  # Fit Random Forest model on training data
  rf_train <- ranger(
    formula         = `# of Cases` ~ Year + hospital_effect + surgery_effect + county_effect,

    data            = trainData,
    num.trees       = 500,
    importance      = 'impurity'
  )

  # Predict on test data
  predictions <- predict(rf_train, data = testData)$predictions

  # Calculate MSE
  results$MSE[i] <- mean((testData$`# of Cases` - predictions)^2)
}

# Calculate the average MSE
mean_mse <- mean(results$MSE)

# Print the average MSE
print(mean_mse)


# Extract MSE for linear regression
linear_mse <- linear_model_cv$results$RMSE^2
```

```
[1] 446.499
[1] 2165.967
```

- The Linear Regression model had an MSE of 2165.967, while the Random Forest model had an MSE of 446.499.

IV. Conclusion

- Now, we will find which model performs better by using Mean Squared Error (MSE). For the Linear Regression model, we found the MSE to be around 2165.967. This is a very high value which indicates that the MSE value is far off from the actual values of the number of cancer cases in California. On the other hand, the MSE of the Random Forest model was found to be around 446.499. In addition, the variability explained by the Random Forest Model is 77.89%. This means that the variability of the Random Forest had a higher variance compared to the Linear Regression model. This indicates that if we have a low MSE and a high variability, then this makes the Random Forest model better than the Linear Regression model. In other words, the Random Forest model fits our data and it is more accurate and reliable to use in our main research question of this project.

V. Rest of the R Code

```r
# Load necessary libraries
library(readr)
library(dplyr)
library(modelr)
library(broom)
library(caret)
library(randomForest)

# Load the data
data <- read_csv("Downloads/Cancer Surgeries CA Hospitals 2013-2022.csv")

# Filter out 'Statewide' entries and drop 'LONGITUDE' and 'LATITUDE' columns
data_filtered <- data %>%
  filter(hospital != "Statewide") %>%
  select(-LONGITUDE, -LATITUDE)

# Compute variance of '# of Cases' and aggregate by 'Hospital', 'Surgery', and 'County'
hospital_variance <- data_filtered %>%
  group_by(hospital) %>%
  summarise(hospital_effect = var(`# of Cases`, na.rm = TRUE))

surgery_variance <- data_filtered %>%
  group_by(Surgery) %>%
  summarise(surgery_effect = var(`# of Cases`, na.rm = TRUE))

county_variance <- data_filtered %>%
  group_by(County) %>%
  summarise(county_effect = var(`# of Cases`, na.rm = TRUE))

# Join these variances back to the main data
data_model <- data_filtered %>%
  left_join(hospital_variance, by = "hospital") %>%
  left_join(surgery_variance, by = "Surgery") %>%
  left_join(county_variance, by = "County")

# Remove rows with NA values
data_model <- na.omit(data_model)

# Fit the linear regression model using the aggregated variables
model <- lm(`# of Cases` ~ Year + hospital_effect + surgery_effect + county_effect, data =
data_model)

# Summarize the model to see coefficients
summary_model <- summary(model)

# Print model summary
print(summary_model)

# Show the residual plots
par(mfrow = c(2,2))
plot
```

VI. Endnotes

- Since this project was done by two students, we worked on every part together and wrote down the project report together. Marwan has focused on the grammar and neatness of the paper while Hasher focused on doing the R scripts and implementing the plots.

VII. References

California Health and Human Services Agency. (2024). *Number of cancer surgeries (volume) performed in California hospitals* [Data set]. Data.gov. Retrieved from https://catalog.data.gov/dataset/number-of-cancer-surgeries-volume-performed-in-california-hospitals-a3f18