Analysis of Space Mission Success

Quang Nguyen and Sean Finn

December 10, 2020

Abstract

This project is a statistical investigation into space missions from 1957 to 2020. We primarily focused on modeling the relationship between space launch outcome and other categorical factors. This work gives us sufficient evidence that there is a difference in the likelihood of space mission success between active and retired rockets; the United States and Russia; and the space race decades - 1960s and 1970s. In particular, our main findings are: 1) active rockets have a higher chance of mission success than retired rockets (OR = 1.696, 95% CI (1.273, 2.302)); 2) Russia is more likely to succeed in conducting space launches than the United States (OR = 1.376, 95% CI (1.099, 1.722)); and 3) there is an increase in space launch success rate from 1960s to 1970s (OR = 3.355, 95% CI (2.512, 4.519)).

Keywords: Space exploration, logistic regression

1 Introduction

Beginning with the launch of Sputnik I in 1957, humans have ventured into outer space and accumulated massive knowledge of the world and our place in it. Missions in Earth orbit and outer space have led to scientific discoveries, vast technological advancements, and a deeper understanding of Earth itself (NASA 2020). In the 63 years since that inaugural launch, humans have attempted to launch rockets into space well over 4000 times. Although the vast majority (92%) of those launches were successful, a number of terrible tragedies have occurred through the decades, resulting in dozens of lives lost (Smith 2018). One particularly jarring example to most Americans was the Space Shuttle Challenger, which exploded 73 seconds after launching on January 28, 1986. The shuttle was carrying seven astronauts, including a high school teacher whose story had garnered the attention of millions of live viewers. The explosion killed all seven people, and caused major shock waves in public opinion about space exploration and NASA, the National Aeronautics and Space Administration.

Throughout the history of rocket development and human spaceflight, launch success has always been a crucial objective. It therefore makes sense that many interested parties, such as government agencies, private corporations, and independent scientists, would be deeply interested in understanding what factors are related to launch success. This was our main objective for this project. While astronauts like Yuri Gagarin and Neil Armstrong set out to explore space, we set out to explore the data they produced through their missions and uncover the statistical differences in space mission success rates.

2 Data

For this project, we used a dataset posted for public use on Kaggle that contains one observation for every recorded launch into Earth orbit or outer space. This encompasses 4324 launches from 1957 through to

present day. These data were originally scraped from NextSpaceflight.com, which records all past and upcoming missions into space from all around the globe. Our data consists of the following attributes: the date of each launch; the company or organization that conducted the launch; the location at which the launch took place; the name of the rocket used; the name of the mission; whether the rocket that was used is still currently active; and whether the launch succeeded or failed. Another variable in the original dataset was cost, which recorded the amount of dollars spent on the mission; however, well over half of the observations did not have values for that variable, and we therefore excluded it from our analysis.

3 Methodology

Our main method for this project is logistic regression, which allows us to model the relationship between the outcome of space missions and other factors, namely, rocket status (active versus retired), country (USA versus Russia), and space race decade (1960s versus 1970s). Logistic regression is appropriate when we have a response variable of binary categorical type, meaning that there are exactly two outcome levels, such as yes-no, pass-fail, or heads-tails. This applies to our case here, since our response variable, space launch outcome, has two categories: success and failure.

A logistic regression model has the form

$$logit(\pi) = log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k,$$

with main components: the log odds of success as the response variable; and a linear predictor $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$, which is a linear combination of the explanatory variables (the X's), with the β's as the regression coefficients to be estimated.

For this particular investigation, each of our models will only consist of one predictor for space mission success, hence they all have the form of a single logistic regression model, which is

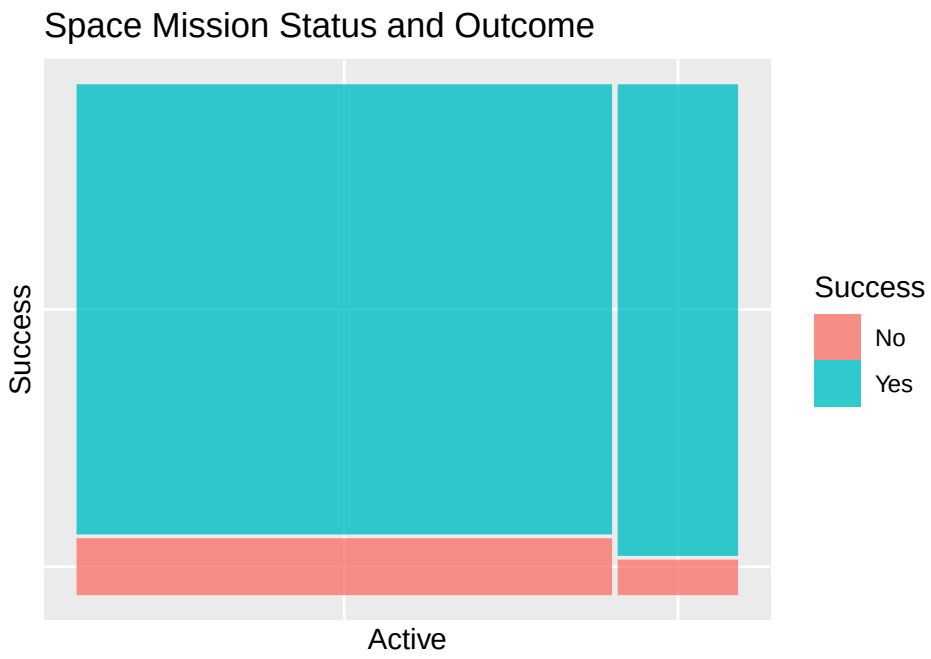
$$logit(\pi) = log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X.$$

As with other regression methods, in logistic regression, we can make inferences about the terms in the model. In particular, we are interested in making hypothesis tests and confidence intervals for the slope term in each of our models. In terms of significance test, we can determine whether the groups or levels in the explanatory variable differ in the chance (odds) of success. The test's null and alternative hypotheses are: $H_o : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$. In addition, the formula for calculating our test statistic is $z = \dfrac{\hat{\beta}_1}{se(\hat{\beta}_1)}$, which can then be used to obtain a two-sided p-value of $P(Z > |z|)$, where Z follows a standard normal distribution. A confidence interval for the slope at significance level α can be computed using the expression $\hat{\beta}_1 \pm z_{(1-\alpha/2)} \cdot se(\hat{\beta}_1)$, where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of a standard normal distribution.

4 Analysis

4.1 Mission Success and Rocket Status

The first question we would like to investigate is: "Is there a difference in the chance of success between rockets with active status and retired rockets?" We start off by looking at a mosaic plot as well as a summary table of our variables of interest.




```
       Active
Success     No    Yes
    No  0.1106 0.0684
    Yes 0.8894 0.9316
```

The proportion of mission success for active rockets is higher than the chance of success for retired shuttles. We then fit a logistic regression model of mission outcome versus rocket status and made inferences about the terms in the model to verify this relationship.

```
             Estimate Std. Error   z value    Pr(>|z|)
(Intercept) 2.0842255 0.05362567 38.866192 0.000000000
ActiveYes   0.5280206 0.15084041  3.500525 0.000464343
```

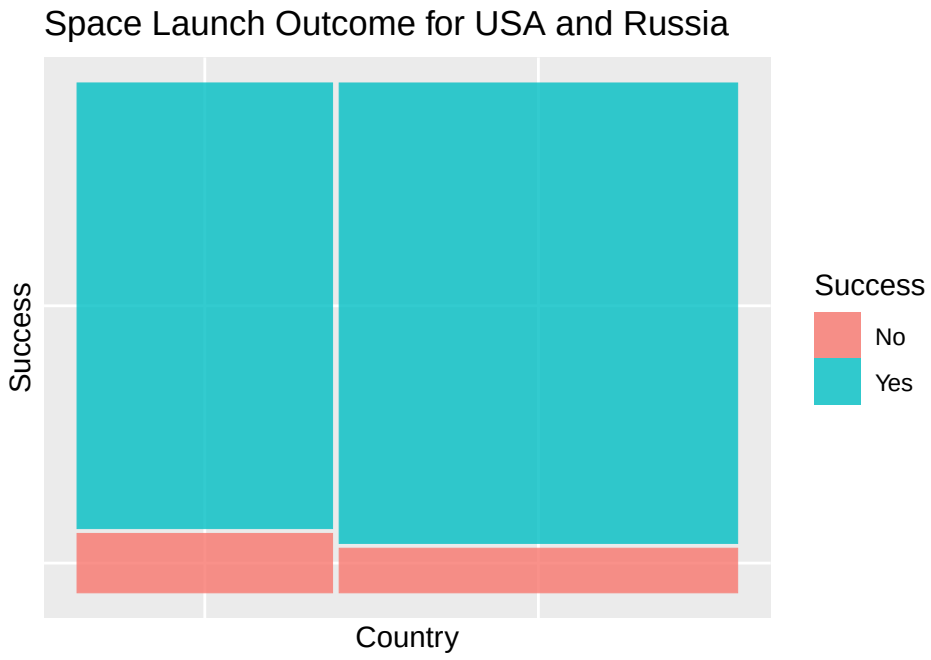
```
ActiveYes     2.5 %    97.5 %
  1.69557   1.27285   2.30176
```

With z = 3.501 and p-value = 0.0005, there is evidence that active rockets and retired rockets differ in the chance of space mission success. The estimated value for slope in our model (0.528) is the log odds ratio of success for the indicated category, active rockets, compared to the reference category, retired rockets. After exponentiating, the odds ratio for the probability of launch success comparing active to retired shuttles is 1.696 (95% CI (1.273, 2.302)), implying that active rockets have 1.696 times greater odds, thus are more likely to succeed, than non-active rockets.

4.2 Mission Success for USA and Russia

Next, we investigated whether there is a difference between the chance of success for space launches conducted by the two giants in the space industry, the United States and Russia. This rivalry began in the late 1950s and early 1960s, as Russia started the competition by launching the first artificial satellite -

Sputnik 1 - in 1957 and sending the first man into space - Yuri Gagarin - in 1961. As a result, the United States and President Kennedy responded by making an ambitious expansion and transformation of the national space program with the primary goal of a landing a man on the moon, which they achieved in 1969. From then on, the race continues, and it is arguably the one of, if not the best rivalry in the history of space exploration. To find out which country is more successful, we begin with some numerical and visual summary of the launches executed by the two nations.




```
       Country
Success    USA Russia
    No  0.1176 0.0883
    Yes 0.8824 0.9117
```

The two countries' probabilities of mission success are very close to one another, but Russia has a small lead over the United States. Indeed, the two nations differ in the likelihood of launch success, as illustrated by our test result (z = 2.789, p-value = 0.005). What we also learn from the test output after back-transforming the estimated slope coefficient is that Russia has 1.376 times greater odds (95% CI (1.099, 1.722)), hence a higher chance, of accomplishing positive space launch outcome than America.

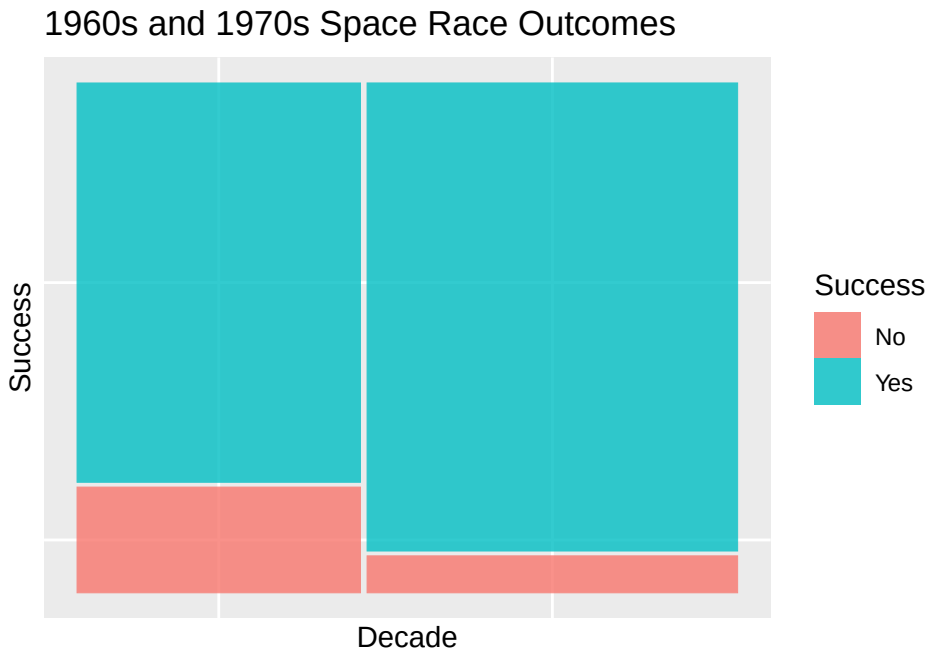
```
              Estimate Std. Error  z value Pr(>|z|)
(Intercept)    2.01575    0.08469 23.80166  0.00000
CountryRussia  0.31928    0.11446  2.78945  0.00528

CountryRussia         2.5 %        97.5 %
      1.37614       1.09885       1.72164
```

4.3 Space Race Decades: 1960s and 1970s

Our last relationship for this investigation concerns launch outcomes for the two space race decades, the 1960s and 1970s. This is arguably the most important era in the history of space exploration. As

mentioned in the previous section, the main highlights of this time period are Yuri Gagarin becoming the first cosmonaut to fly into outer space, and the peak of the space race, which is the Apollo 11 mission that landed the first humans on the Moon. In addition to those historical events, this era meant a lot to the future of astronomy and space science, as it paved the path for more space innovation and exploration in the following years. Similar to the previous two analyses, we explored how successful the space missions were in the 1960s and 1970s by examining some summary figures for our variables.




```
       Decade
Success   1960s   1970s
    No  0.20930 0.07312
    Yes 0.79070 0.92688
```

From the plot and statistics above, it is clear that the 1970s has a higher success rate (92.7%) than the 1960s (79.1%). The test results from the table of coefficients confirm that the chance of success for space missions is different for the two space race decades. In particular, there is an increase in space launch success rate from the 1960s to 1970s, as demonstrated by an odds ratio of 3.355 (95% CI (2.512, 4.519)).

```
            Estimate Std. Error   z value     Pr(>|z|)
(Intercept) 1.329136 0.08835625 15.042920 3.842221e-51
Decade1970s 1.210549 0.14962088  8.090775 5.928609e-16
```

```
Decade1970s       2.5 %      97.5 %
    3.35533     2.51205     4.51942
```

5 Model Fit

After fitting and analyzing our models, we are interested in how well our models fit their respective set of data. To measure the goodness-of-fit for our models, we use the R^2 statistic for logistic regression. R^2

can be computed as

$$R^2 = 1 - \frac{Residual\ Deviance}{Null\ Deviance},$$

where the residual deviance is the deviance of our fitted (full) model, showing how well the response is predicted by the model when the predictor is included. On the other hand, the null deviance represents how effectively the logit is predicted by a reduced model with only the intercept.

There are similarities and differences between this R^2 for logistic regression and the linear regression version of R^2. In relation to the coefficient of determination, the residual deviance is equivalent to the error sum of squares, measuring the differences between the observed and predicted values for the response variable; whereas the null deviance is similar to the total sum of squares, which describes the differences between the observed dependent variable and its mean. Unlike linear regression's R^2, the R^2 for our logit model does not indicate the percentage of the variability in the dependent variable that the independent variable explains collectively, but instead it is a ratio of how perfect or bad the fit is between the model and the data. The statistic is also not related to any correlation coefficient, hence is not a measure of the relationship strength between the variables.

Rocket Status	Country	Decade
0.00477	0.00345	0.05083

We calculated the R^2 for our three logistic models, which are displayed in the table above. As we can see, the fits are not great here, as all R^2 values are closer to 0 rather than to a perfect fit of 1. The logistic regression model comparing space launch success for the two space race decades did the least worst of fitting its corresponding set of data compared to the other two models. This could be something to improve in the future, as we could try to find the best methods or subset of variables for predicting space mission success that give us the best fit.

6 Conclusion

In summary, we discovered multiple statistically meaningful relationships between rocket launch success and other factors. Unsurprisingly, rockets that are currently active have 1.696 times greater odds of success than those that are no longer active. This makes sense, as rockets are retired when technology improves or when, for example, they explode. Additionally, we found that space shuttles launched in the 1970s had 3.355 times greater odds of success than those launched in the 1960s. These results comport with common sense, although they would not be useful models for predicting the outcome for space missions in the future, as all future means of transportation used will be active by definition and launches will not take place during either space race decade.

Another interesting finding, and one with potential value for predicting future launch success, is that Russia has 1.376 times greater odds of success than the USA. This may be due to any number of factors that are country-specific. Possible explanations include funding, scientific or technological advances, or level of agency oversight. Although we do not currently have data with which we can explore those explanations, this finding does open the door to the interesting question of why Russia has a higher mission success chance than America. This is a possible avenue for future research into this topic.

Overall, it is important to keep in mind that these models provide small, incomplete pictures of the factors influencing the result of space launches. Each model analyzes only one variable and the goodness-of-fit measures reveal that they do not fit the data well. In order to build a more complete picture, many more details about each launch would be necessary. Other future statistical analyses with these data may include multivariate regression, principal component analysis, neural networks, or time series analysis.

7 References

Dunn, P. K., and G. K. Smyth. 2018. *Generalized Linear Models with Examples in R.* Springer Texts in Statistics. Springer New York.

Fox, J. 2015. *Applied Regression Analysis and Generalized Linear Models.* SAGE Publications.

Merola, C. (Kaggle username: Agirlcoding). 2020. "All Space Missions from 1957." https://www.kaggle.com/agirlcoding/all-space-missions-from-1957.

NASA. 2020. "NASA History." National Aeronautics and Space Administration. https://www.nasa.gov/topics/history.

Smith, M. 2018. "NASA's 2018 Day of Remembrance Honors Fallen Astronauts." SpacePolicyOnline.com. https://spacepolicyonline.com/news/nasas-2018-day-of-remembrance-honors-fallen-astronauts.

8 Credit

Quang performed exploratory data analysis and researched and wrote descriptions of the logistic regression model. He also wrote the code to perform logistic regression in R and calculate the goodness-of-fit measure. Quang wrote the abstract, methodology, analysis, and model fit sections of the paper. Sean found the dataset, performed exploratory data analysis, wrote some of the code for data visualizations and researched goodness of fit measurement options. He wrote the introduction, data, and conclusion sections of the paper. Together we decided the model selection, wrote code for data visualization, and created the project presentation and report in R Markdown.

9 Appendix

The R code for this document can be found below. In addition, all of the materials related to this project are publicly available on GitHub at https://github.com/qntkhvn/spacemissions.

```
# Packages
library(tidyverse)
library(mosaic)
library(ggmosaic)

# Data
data <- "https://raw.githubusercontent.com/qntkhvn/spacemissions/main/spacemission.csv"
space <- read.csv(data)
sp <- space[,3:ncol(space)] %>%
  separate(Datum, into = c("Weekday", "Month", "Day", "Year")) %>%
  mutate(Active = as.factor(ifelse(Status.Rocket == "StatusActive", "Yes", "No")),
         Success = as.factor(ifelse(Status.Mission == "Success", "Yes", "No"))) %>%
  rename(Company = Company.Name)

# Model 1: Rocket Status
## Mosaic plot
ct <- count(sp, Active, Success)
ggplot(data = ct) +
  geom_mosaic(aes(weight = n, x = product(Active), fill = Success)) +
  xlab("Active") + ylab("Success") +
  ggtitle("Space Mission Status and Outcome")
## 2x2 table
tally(Success ~ Active, format = "proportion", data = sp) %>% round(4)
## Model
mod <- glm(Success ~ Active, family = binomial, data = sp)
summary(mod)$coef %>% round(5)
c(exp(coef(mod)[2]), exp(confint(mod)[2,])

# Model 2: USA vs Russia
## Transform the data to get the countries
sp <- sp %>%
  mutate(ID = paste("id", 1:nrow(sp), sep = "")) %>%
  separate(Location, into = c("str1", "str2", "str3", "str4"), sep = ", ")
usa <- sp %>%
  filter(str4 == "USA")
rus <- sp %>%
  filter(str3 %in% c("Russia", "Kazakhstan") | str2 == "Russia")
sp <- sp %>%
  mutate(Country = ifelse(ID %in% usa$ID, "USA",
                          ifelse(ID %in% rus$ID, "Russia", "Other")))
usrus <- sp %>%
  filter(Country %in% c("USA", "Russia"))
usrus$Country <- factor(usrus$Country, levels = c("USA", "Russia"))
# Summary plot and table
ct <- count(usrus, Country, Success)
```

```r
ggplot(data = ct) +
  geom_mosaic(aes(weight = n, x = product(Country), fill = Success)) +
  xlab("Country") + ylab("Success") +
  ggtitle("Space Mission Outcome for USA and Russia")
tally(Success ~ Country, format = "proportion", data = usrus) %>% round(4)
# Model
usrus_mod <- glm(Success ~ Country, family = binomial, data = usrus)
coef(summary(usrus_mod)) %>% round(5)
c(exp(coef(usrus_mod)[2]), exp(confint(usrus_mod)[2,])

# Model 3: Space Race Decades
## Transform the data to get the decades
sprace <- sp %>%
  filter(as.numeric(Year) %in% 1960:1979) %>%
  mutate(Decade = ifelse(Year %in% 1960:1969, "1960s", "1970s"))
## Plot and tables
ct <- count(sprace, Decade, Success)
ggplot(data = ct) +
  geom_mosaic(aes(weight = n, x = product(Decade), fill = Success)) +
  xlab("Decade") + ylab("Success") +
  ggtitle("1960s and 1970s Space Race Outcomes")
tally(Success ~ Decade, format = "proportion", data = sprace) %>% round(4)
## Model
dec_mod <- glm(Success ~ Decade, family = binomial, data = sprace)
coef(summary(dec_mod))
c(exp(coef(dec_mod)[2]), exp(confint(dec_mod)[2,])

## Goodness-of-fit: R^2
Rsq_log_reg <- function(model) {
  1 - model$deviance / model$null.deviance
}
Rsquared <- rbind.data.frame(c(Rsq_log_reg(mod),
                               Rsq_log_reg(usrus_mod),
                               Rsq_log_reg(dec_mod)))
colnames(Rsquared) <-  c("Rocket Status", "Country", "Decade")
Rsquared %>% round(5)
```