

Analysis of Space Mission Success

Quang Nguyen and Sean Finn

December 04, 2020

Abstract

This project is a statistical investigation into space missions from 1957 to 2020. We primarily focused on modeling the relationship between mission 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 greater chance of success than retired rockets (OR = 1.696, 95% CI (1.273, 2.302)), 2) Russia is more likely to have successful space missions than USA (OR = 1.376, 95% CI (1.099, 1.722)), and 3) there's an increase in success rate from 1960s to 1970s (OR = 3.355, 95% CI (2.512, 4.519)).

Keywords: Space exploration, logistic regression

1 Introduction

2 Data

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, mission outcome, has two categories: success and failure.

A logistic regression model has the form

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

with main components: the log odds of success as the response variable; and a linear predictor $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$, 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

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

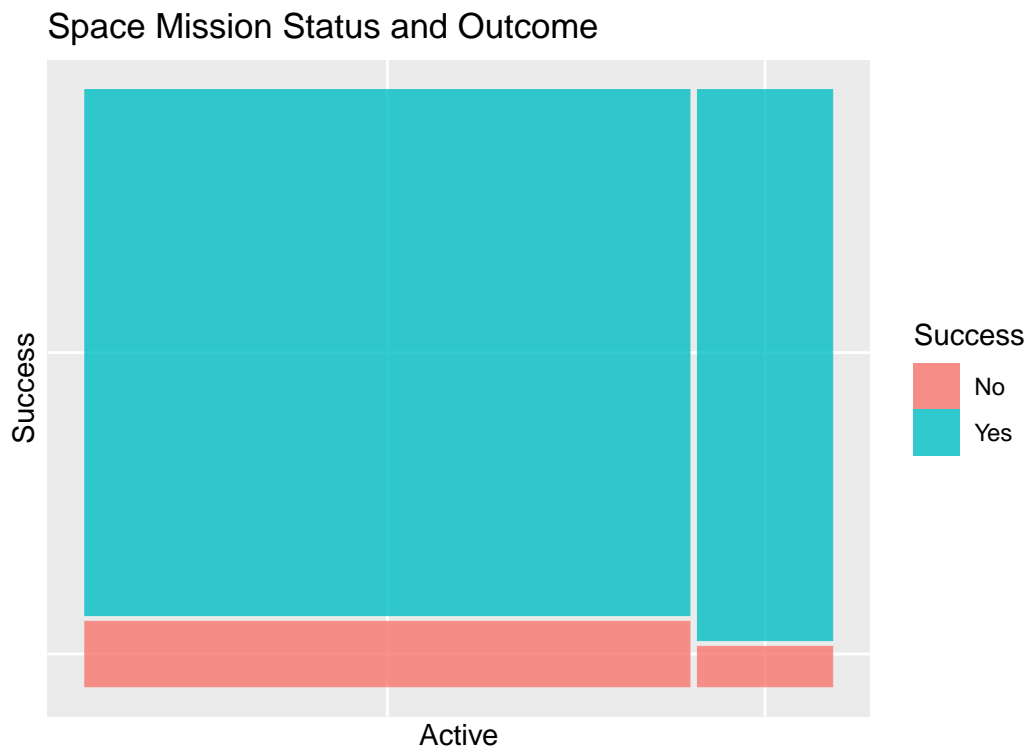
As with other regression methods, in logistic regression, we can make inferences about the terms in the model. In particular, we are interested in making hypothesis tests and confidence intervals for the slope term in each of our models. In terms of significance test, we can determine whether the groups or levels in the explanatory variable differ in 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 $z = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)}$, 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 $\hat{\beta}_1 \pm z_{(1-\alpha/2)}s(\hat{\beta}_1)$.

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

As we can see, the proportion of mission success for active rockets is higher than the chance of success for retired rockets. We then fit a logistic regression model of mission outcome versus rocket status and make inferences about the terms in our model to verify this relationship.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.08423	0.05363	38.86619	0.00000
ActiveYes	0.52802	0.15084	3.50052	0.00046

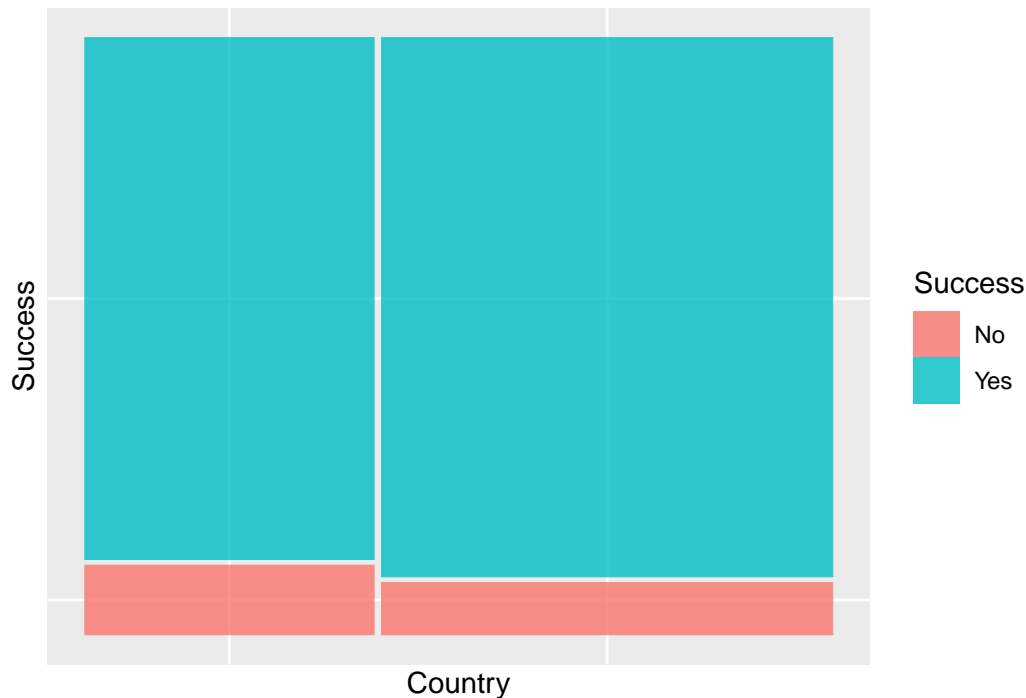
ActiveYes	2.5 %	97.5 %
1.69557	1.27285	2.30176

With $z = 3.501$ and $p\text{-value} = 0.0005$, we have evidence that active rockets and retired rockets differ in chance of space mission success. If we focus on the estimated value for slope term of our model (0.528), it 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 are going to find out whether there is a difference between the chance of success for space missions 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 to their national space program, with the goal of a moon landing, 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.

Space Mission Outcome for USA and Russia



Country		
Success	USA	Russia
No	0.1176	0.0883
Yes	0.8824	0.9117

The probability of mission success between the two countries are very close to one another, but Russia has a small lead over the United States. Indeed, the two nations differ in chance of mission success, as illustrated by our test result ($z = 2.789$, $p\text{-value} = 0.005$). 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 having successful space missions 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

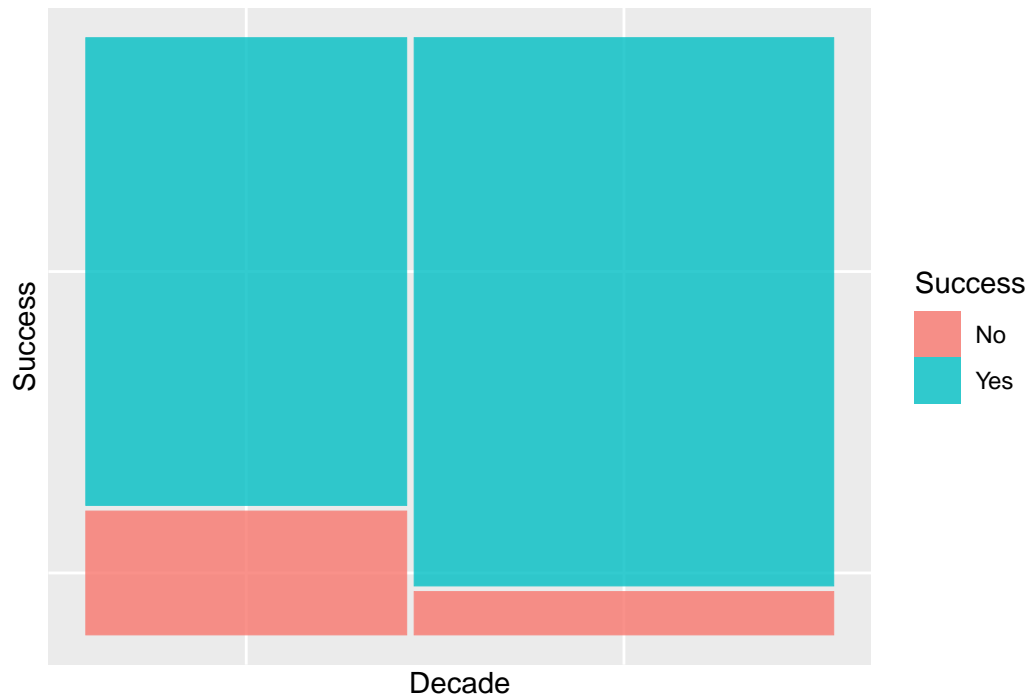
CountryRussia	2.5 %	97.5 %
1.37614	1.09885	1.72164

4.3 Space Race Decades: 1960s and 1970s

Our last relationship for this investigation is about mission outcome for the two space race decades, 1960s and 1970s, which 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 means 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, to begin our analysis on how successful the space missions were in the 1960s and 1970s, we first take a quick look at some figures of summary for our variables.

1960s and 1970s Space Race Outcomes



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

From the mosaic 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 outputted table of coefficients confirm that the chance of success for space missions is different for the 1960s and 1970s. 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

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

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(s) are included. On the other hand, the null deviance shows how well the logit is predicted by a reduced model with only the intercept (the grand mean).

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 , logistic regression's R^2 does not indicate the percentage of the variability in the dependent variable that the independent variable(s) explain collectively, but instead it is a ratio of how perfect or worst 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
1	0.00477	0.00345	0.05083	

We then calculated the R^2 for each of our three models. As we can see, the fits do not look that great here, as all of our R^2 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 than 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 predictors for space mission success that give us the best fit.

6 Conclusion and Discussion

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>.

8 Credit

9 Appendix

The R code for this manuscript 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)

# 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,]))

# 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)
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,]))

# Space Race

## 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:  $R^2$ 
Rsqr_log_reg <- function(model) {
  1 - model$deviance / model$null.deviance
}

Rsquared <- rbind.data.frame(c(Rsqr_log_reg(mod),
                              Rsqr_log_reg(usrus_mod),
                              Rsqr_log_reg(dec_mod)))
colnames(Rsquared) <- c("Rocket Status", "Country", "Decade")

```



```
Rsquared %>% round(5)
```