

# Analyzing Flight Dynamics: A Factorial Design Study Using Paper Airplanes

Kacie Chong

2025-05-15

## Introduction:

Understanding how different design choices can affect performance lies at the heart of both science and engineering. A fully loaded passenger aircraft must consider the added weight of luggage, passengers, fuel, and various cargo, all of which significantly impact its balance, stability, and flight performance. In fact, designers and engineers need to know about the distribution of aircraft loads because the plane configurations must be able to withstand them (Kundu, 2010). Engineers must carefully calculate how this weight is distributed to make sure the plane is safe for traveling, and “weight prediction and control are part and parcel of every phase in the design and development of every aircraft” (Torenbeek, 1982).

While the scale is very different from a commercial airplane, the underlying physics for paper airplanes is the same. A simple hands-on experiments in my own bedroom can reflect real world challenges actually faced in aerospace design and engineering. Adding something as small as a paperclip to different parts of a paper airplane can shift the center of mass and dramatically affect how the plane glides or nosedives.

As you can see, the concept of weight remains a critical study of aircraft design, but we must not only study how much an aircraft weighs, but also where that weight is located. This is why “more sophisticated weights methods estimate the weight of the various components of the aircraft and then sum for the total empty weight” (Raymer, 2012). This highlights that beyond total weight, the distribution of weight is important. Experiments like paper airplane studies make it easier to explore fundamental principles such as force, motion, and stability in a fun, low-stakes way. While it may seem like child’s play, these simple experiments offer valuable insights into flight mechanics, showing how even small changes in weight distribution can significantly affect flight distance.

In traditional Randomized Complete Block Design studies like the one conducted in Lab 5 of this course, each treatment level, “paperclip at the nose”, “paperclip at the rear”, and “no paperclip”, is tested in isolation with one or zero added clips. However, this does not allow us to understand how combinations of paperclip placements influence outcomes. To get a more complete picture, we need to explore how multiple factors interact, especially when those factors most likely do not act independently. A full factorial approach will give us insights into how combinations of factors may enhance or counteract each other.

This project builds on previous work completed in Lab 5 by conducting a full  $2^3$  factorial experiment, which allows us to evaluate all combinations of three binary factors: clip placement at the nose, middle, and rear of the plane. Each factor is tested at two levels, with or without a clip, resulting in 8 unique treatment combinations. The combination of hands-on data collection with factorial modeling will give us a glimpse into both the science of flight and the statistical tools needed to interpret experimental results. These kinds of small-scale studies are great for building intuition while also giving you a chance to practice essential statistical skills like hypotheses testing, controlling variables, and discovering how different factors interact. I hypothesized that attaching a paperclip to the nose would have a significant effect on flight distance because adding weight to the front of the paper airplane can cause it to tilt downward or change its flight path. All in all, the overall goal of this study is to determine how the location of paperclips affects flight distance and whether there are interaction effects between clip placements.

## Methods:

### Data Collection and Experimental Setup

The data for this experiment was collected by throwing paper airplanes with varying placements of paperclips to test their effect on flight distance. The required materials was a 8.5" x 11" sheet of printer paper, 3 paper clips, and a measuring tape. A single paper airplane design was folded following the directions from HGTV.com's "How to Make a Paper Airplane". The experiment used a full factorial  $2^3$  design, meaning three factors were tested at two levels each where "yes" means a standard metal paperclip was attached to the corresponding position and "no" means the corresponding position was not clipped. In other words, three independent factors were tested, each at two levels as shown below:

#### 1. Factor A: Paperclip on the Nose

- Level: Yes (1) – A standard metal paperclip is attached to the nose of the object.
- Level: No (0) – No paperclip is attached to the nose.

#### 2. Factor B: Paperclip on the Middle

- Level: Yes (1) – A paperclip is attached to the middle of the object.
- Level: No (0) – No paperclip is attached to the middle.

#### 3. Factor C: Paperclip on the Rear

- Level: Yes (1) – A paperclip is attached to the rear.
- Level: No (0) – No paperclip is attached to the rear.

This setup resulted in 8 unique treatment combinations, representing all possible combinations of the three paperclip placements.

To reduce variability in flight performance unrelated to paperclip placement, the same paper airplane was used for all throws. This is because folding airplanes can introduce subtle differences in shape, which could confound the results. By using a single airplane, the experiment focused specifically on how the location of the paperclips affected flight distance. The airplane was checked regularly for damage, and, to maintain precise placement of the clips, the airplane was marked with spots using a pencil indicating the nose, middle, and rear positions. Paperclips of the same size and weight were attached according to the assigned treatment for each throw. The throws were conducted indoors in my bedroom to minimize outside variability such as wind and weather conditions. I aimed to keep the throwing force, angle, and stance consistent throughout to minimize variability from human factors. Flight distances were measured in centimeters from the starting point to the point where the airplane first landed, using a tape measure laid flat on the floor. Even if the plane veered sideways or backward, I only measured forward progress along the tape measure, with sideways or backward travel recorded as 0. Sliding after landing was not counted toward flight distance. These distances were recorded immediately after each throw.

```
library(knitr)
knitr::include_graphics("/Users/kaciechong/Desktop/pic.png")
```

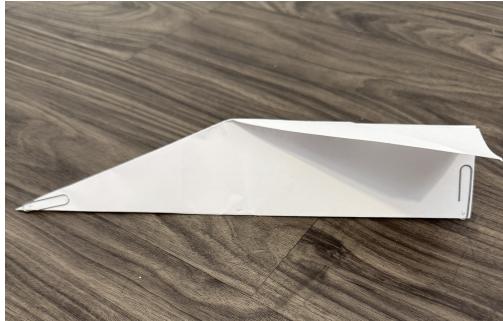


Figure 1: Example Airplane With Clips On Nose and Rear

```
knitr::include_graphics("/Users/kaciechong/Desktop/pic2.png")
```



Figure 2: Data Collection Setup

## Sample Size Calculation

Before conducting the full experiment, a pilot study was performed to estimate the number of replicates needed per treatment condition in order to achieve sufficient statistical power. For context, pilot studies are a common and effective way to assess baseline variability and effect sizes using a smaller data set before committing to the full data collection. In this pilot, I conducted 3 replicates for each of the 8 treatment combinations, resulting in 24 total throws. The order of these throws was randomized using R's `sample()` function to prevent potential bias from order. Then, I fit a linear model using R's `lm()` function to estimate the mean effect sizes (`beta_mean`) and standard errors (`beta_se`) for the three clip position factors. These

estimates were then used as inputs in a power analysis using the `power_factorial_23()` function. To explore the impact of different levels of variability on power, the analysis considered three scenarios:

- Baseline variability (using `beta_se` from the pilot model)
- Low variability ( $\text{beta\_se} \times 0.75$ )
- High variability ( $\text{beta\_se} \times 1.5$ )

To summarize the power analysis results, I created a line plot illustrating how statistical power varied with the number of replicates under three different variability scenarios. The visualization helped identify that at least 5 replicates per treatment condition were necessary to achieve sufficient power of 80% for detecting effects. Based on this, I conducted the main experiment using 40 total throws, 5 for each of the 8 treatment combinations. The throwing order was once again randomized using R's `sample()` function.

## Hypotheses

- Null hypothesis: The mean flight distances are the same for all combinations of paperclip placements, meaning there are no effect of nose, middle, or rear clips, nor their interactions.
- Alternative hypothesis: At least one combination of paperclip placements results in a different mean flight distance, so at least one factor or interaction affects flight distance.

## Statistical Methods

I analyzed the data using a full factorial linear model, fitted with the `lm()` function. The design followed a  $2^3$  full factorial structure, with three binary (yes/no) factors representing paperclip placement at the nose, middle, and rear of the airplane. The model included all possible interaction terms, allowing us to assess not only the individual effects of each clip placement, but also whether specific combinations of placements influence flight distance differently than expected from their individual effects alone. The `lm()` output also provides estimated effect sizes, giving a quantitative sense of how much each factor and interaction contributes to changes in flight distance.

Because factorial linear models rely on several key assumptions, I carefully evaluated diagnostic plots and statistical tests to verify that these assumptions were reasonably satisfied before interpreting the results:

- Independence of observations: Each throw should be independent of the others, so I fully randomized the order of throws using R's `sample()` function. I will also examine the residuals vs. order of data collection plot.
- Normality of residuals: The residuals from the model should be approximately normally distributed because small sources of error like slight differences in how the planes are thrown or subtle changes in air movement are mostly random and tend to cancel each other out. I will examine a residual histogram and Q-Q plot and perform a Shapiro-Wilk test to confirm.
- Equal variances across all fitted values: The variance across different treatment groups should be roughly equal because the only differences between groups are the positions of the paperclips which were controlled and deliberately varied, and the order of throws was randomized. I also ensured that the paper airplane remained in good condition throughout the experiment, so wear and tear from repeated throws did not introduce extreme additional variability. The amount of random error or variability in each group should be similar. I will also observe the spread of a plot of residuals vs. fitted values to confirm.

## Technical Issues

- Throwing consistency: I made a conscious effort to use a consistent technique for every throw using the same stance, arm motion, and release angle. However, natural human variation is unavoidable. Considering I was performing 24 throws for the pilot study and 40 throws for the actual study all in one day, my physical energy may have decreased throughout the experiment.
- Airplane wear and tear: Using the same paper airplane across all trials helped reduce design variability. However, repeated handling and impact with the ground may have led to subtle changes in the plane's structure such as minor crumpling near the nose, especially in later throws. Notably, I used a different airplane for the pilot study and the actual study.
- Paperclip placement precision: To minimize placement error, I marked each of the three possible clip positions on the airplane before starting the experiment. This helped ensure the paperclips were attached in the same place every time, but some small shifts may still have occurred during handling.
- Environmental factors: The experiment was conducted indoors to eliminate wind and weather as confounding variables. However, room air currents or nearby movement may have had slight, unpredictable effects on some flights.
- Measurement error: Flight distances were measured using a tape measure from the throwing point to where the airplane first touched the ground. Human error in reading or recording the distances may have occurred, adding some variability to the data.

Overall, none of these issues appeared to systematically bias the results, but they are worth noting as sources of potential variability.

## Results:

### Sample Size Calculations

#### Conducting the Pilot Study

To determine the appropriate number of replicates for the full factorial experiment, I performed a pilot study using a small number of trials ( $n = 3$ ) across all treatment combinations. This smaller-scale trial helped me estimate the size of the effects and the variability in my measurements. By analyzing the pilot data, I can perform a sample size calculation to determine how many replicates I would need to confidently detect real differences in flight distance.

The order of the throws for the pilot study was randomized using R's `sample()` function:

```
## Define 8 treatment combinations for the 3 factors
treatments <- data.frame(
  Nose = c("yes", "yes", "yes", "yes", "no", "no", "no"),
  Middle = c("yes", "yes", "no", "no", "yes", "yes", "no"),
  Rear = c("yes", "no", "yes", "no", "yes", "no", "yes")
)

## Repeat each treatment combination 3 times
replicated_treatments <- treatments[rep(1:8, each = 3), ]

## Randomize the order of throws
set.seed(123)
randomized_order <- replicated_treatments[sample(1:nrow(replicated_treatments)), ]
```

```

## Rename columns to ensure clear headers for the table output
colnames(randomized_order) <- c("Nose", "Middle", "Rear")

## Remove row names so numbers don't appear in the table
rownames(randomized_order) <- NULL

kable(randomized_order,
      caption = "Randomized Treatment Order for Pilot Study")

```

Table 1: Randomized Treatment Order for Pilot Study

|  | Nose | Middle | Rear |
|--|------|--------|------|
|  | no   | yes    | yes  |
|  | no   | no     | yes  |
|  | no   | yes    | yes  |
|  | yes  | yes    | yes  |
|  | yes  | no     | no   |
|  | no   | yes    | no   |
|  | yes  | no     | no   |
|  | yes  | yes    | no   |
|  | yes  | yes    | no   |
|  | no   | no     | no   |
|  | yes  | yes    | no   |
|  | yes  | no     | yes  |
|  | no   | no     | yes  |
|  | no   | yes    | no   |
|  | no   | no     | yes  |
|  | no   | yes    | yes  |
|  | yes  | yes    | yes  |
|  | no   | yes    | no   |
|  | yes  | no     | yes  |
|  | yes  | no     | no   |
|  | no   | no     | no   |
|  | yes  | no     | yes  |
|  | yes  | yes    | yes  |

I created a data set of the resulting distances after performing the pilot study. Then, I fitted a linear model using R's lm() function to estimate the effect sizes and standard errors for each paperclip placement factor and their interactions. These estimates, specifically the coefficients and their standard errors, provided a quantitative measure of how each factor influenced flight distance and how much variability was present.

```

# Create dataset with flight distances and paperclip placements
pilot_data <- data.frame(
  distance = c(150, 180, 160, 145, 240, 170, 245, 225, 220, 200, 230, 165, 175, 195, 185,
              155, 145, 210, 180, 215, 235, 190, 225, 160),
  nose = c("no", "no", "no", "yes", "yes", "no", "yes", "yes", "no", "yes", "yes",
          "no", "no", "no", "no", "yes", "no", "yes", "yes", "no", "yes", "yes"),
  middle = c("yes", "no", "yes", "yes", "no", "yes", "no", "yes", "yes", "no", "yes",
             "no", "no", "yes", "no", "yes", "yes", "no", "yes", "no", "yes", "no",
             "no", "no", "yes", "no", "yes", "yes", "no", "no", "no", "no", "yes"),
  rear = c("yes", "yes", "yes", "yes", "no", "no", "no", "no", "no", "no", "no",
           "no", "no", "yes", "no", "yes", "yes", "no", "no", "no", "no", "yes",
           "no", "no", "yes", "no", "yes", "yes", "no", "no", "no", "no", "yes")

```

```

    "yes", "no", "no", "yes", "yes", "yes", "no", "yes", "no", "no", "yes", "yes")
)

# Convert placement variables to factors
pilot_data$nose <- factor(pilot_data$nose)
pilot_data$middle <- factor(pilot_data$middle)
pilot_data$rear <- factor(pilot_data$rear)

# Fit a linear regression model with all main effects and interactions
model_fit <- lm(distance ~ nose * middle * rear, data = pilot_data)

# Extract model coefficients and format output for display
coefficients_summary <- summary(model_fit)$coefficients
rounded_coeffs <- signif(coefficients_summary, digits = 4)
rounded_coeffs[,] <- as.character(rounded_coeffs[,])

knitr::kable(rounded_coeffs, caption = "Regression Coefficients from Pilot Study")

```

Table 2: Regression Coefficients from Pilot Study

|                           | Estimate | Std. Error | t value  | Pr(> t )  |
|---------------------------|----------|------------|----------|-----------|
| (Intercept)               | 195      | 10.32      | 18.89    | 2.309e-12 |
| noseyes                   | 45       | 14.6       | 3.082    | 0.007145  |
| middleyes                 | -16.67   | 14.6       | -1.141   | 0.2705    |
| rearyes                   | -25      | 14.6       | -1.712   | 0.1062    |
| noseyes:middleyes         | 1.667    | 20.65      | 0.08071  | 0.9367    |
| noseyes:rearyes           | -13.33   | 20.65      | -0.6457  | 0.5276    |
| middleyes:rearyes         | -1.667   | 20.65      | -0.08071 | 0.9367    |
| noseyes:middleyes:rearyes | -13.33   | 29.2       | -0.4566  | 0.6541    |

### Power Analysis for Sample Size Determination

Next, I used these estimated effect sizes (`beta_mean`) and standard errors (`beta_se`) as inputs in `power_factorial_23()`. This allowed me to calculate the statistical power for different numbers of replicates under varying assumptions about variability: baseline, lower, and higher variability. I wanted to determine how many throws per treatment combination would be needed in the main study to reliably detect the effects observed in the pilot data with 80% power. Following my TA's instructions, to simplify and standardize the sample size calculation, the `beta_mean` and `beta_se` vectors were updated with rounded values. The new `beta_mean` values were chosen to reflect uniform effect sizes, set to 10 for non-intercept terms, to represent a consistent, moderate impact across predictors. Similarly, the `beta_se` values were rounded to the nearest practical value to generalize variability estimates while maintaining reasonable approximation of the original standard errors.

```

set.seed(123)

## Load custom power function
source("/Users/kaciechong/Desktop/power_factorial_23 (1).R")

## Mean effect sizes for each term:
### Intercept, nose, middle, rear, nose:middle, nose:rear, middle:rear, nose:middle:rear

```

```

# Original means: beta_mean <- c(195, 45, -16.67, -25, 1.667, -13.33, -1.667, -13.33)
# Rounded means for simplified sample size calculation:
beta_mean <- c(200, 10, 10, 10, 10, 10, 10, 10)

## Standard errors for each term
# Original standard errors: beta_se <- c(10.32, 14.6, 14.6, 14.6, 20.65, 20.65, 20.65, 29.2)
# Rounded standard errors for simplified sample size calculation:
beta_se <- c(12, 12, 12, 12, 20, 20, 20, 20)

# Create lower and higher variability versions of standard errors
beta_se_low <- beta_se * 0.75
beta_se_high <- beta_se * 1.5

# Define sample sizes to test (number of replicates per condition)
reps <- seq(2, 15, by = 1)

# Initialize vectors to store power results
power_low <- numeric(length(reps))
power_baseline <- numeric(length(reps))
power_high <- numeric(length(reps))

# Loop over different sample sizes and calculate power under three variability scenarios
for (i in seq_along(reps)) {
  power_low[i] <- power_factorial_23(beta_mean, beta_se_low, reps[i])
  power_baseline[i] <- power_factorial_23(beta_mean, beta_se, reps[i])
  power_high[i] <- power_factorial_23(beta_mean, beta_se_high, reps[i])
}

# Combine all results into one data frame
sample_sizes <- data.frame(
  power = c(power_low, power_baseline, power_high),
  beta_se = factor(rep(c("low", "baseline", "high"), each = length(reps))),
  reps = rep(reps, times = 3)
)

```

## Visualizing the Power Analysis

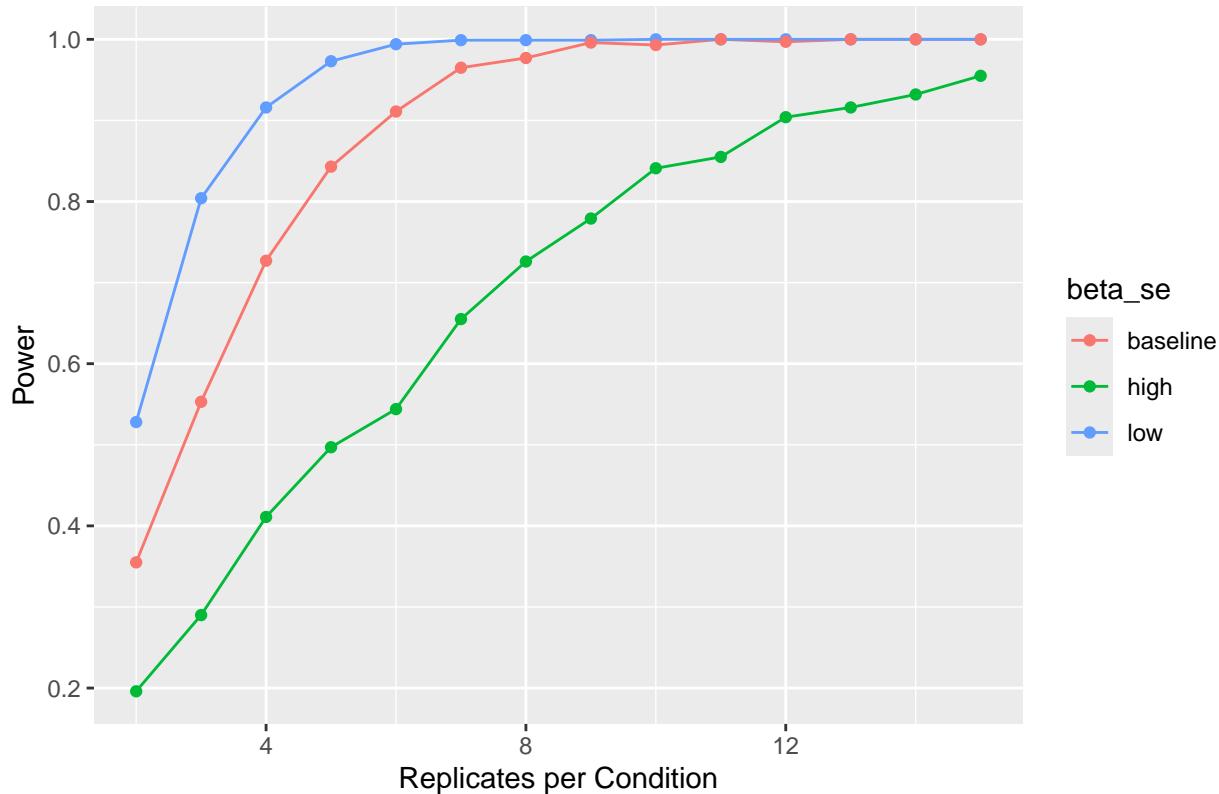
```

library(ggplot2)

ggplot(sample_sizes, aes(x = reps, y = power, color = beta_se)) +
  geom_line() +
  geom_point() +
  labs(title = "Power vs Sample Size",
       x = "Replicates per Condition",
       y = "Power")

```

## Power vs Sample Size



This plot clearly displays how many replicates you need to reach a desirable power level of 80%. As you can see, five replicates per treatment combination appear sufficient to achieve adequate statistical power under baseline variability.

## Performing the Actual Experiment

I will repeat the same randomization process using R's `sample()` function, which I used in the pilot study, for the actual study. However, there will be more replicates per treatment, five instead of three based on my sample size calculations.

```
## Define 8 treatment combinations for the 3 factors: Nose, Middle, Rear
treatments <- data.frame(
  Nose = c("yes", "yes", "yes", "yes", "no", "no", "no", "no"),
  Middle = c("yes", "yes", "no", "no", "yes", "yes", "no", "no"),
  Rear = c("yes", "no", "yes", "no", "yes", "no", "yes", "no")
)

## Repeat each treatment combination 5 times
replicated_treatments <- treatments[rep(1:nrow(treatments), each = 5), ]

## Randomize the order of throws
set.seed(123)
randomized_order <- replicated_treatments[sample(1:nrow(replicated_treatments)), ]

## Rename columns to ensure clear headers for the table output
colnames(randomized_order) <- c("Nose", "Middle", "Rear")
```

```

## Remove row names so numbers don't appear in the table
rownames(randomized_order) <- NULL

kable(randomized_order,
      caption = "Randomized Treatment Order for Actual Study")

```

Table 3: Randomized Treatment Order for Actual Study

| Nose | Middle | Rear |
|------|--------|------|
| no   | no     | yes  |
| yes  | no     | yes  |
| yes  | no     | yes  |
| yes  | yes    | yes  |
| no   | no     | no   |
| no   | yes    | yes  |
| no   | yes    | no   |
| no   | yes    | no   |
| no   | no     | yes  |
| yes  | yes    | yes  |
| yes  | no     | no   |
| no   | no     | yes  |
| no   | no     | yes  |
| no   | yes    | no   |
| yes  | yes    | no   |
| no   | no     | no   |
| yes  | yes    | no   |
| yes  | yes    | no   |
| yes  | yes    | no   |
| no   | no     | yes  |
| no   | yes    | no   |
| yes  | no     | no   |
| yes  | yes    | yes  |
| no   | no     | no   |
| no   | yes    | yes  |
| no   | no     | no   |
| no   | yes    | yes  |
| yes  | no     | yes  |
| yes  | yes    | yes  |
| yes  | yes    | no   |
| yes  | no     | no   |
| no   | yes    | yes  |

After randomizing the treatment order, I conducted all the paper airplane throws according to this sequence.

Then, I measured and recorded the flight distances for each throw, compiling these data into a data frame for further analysis.

```
# Create dataset with flight distances and paperclip placements
actual_data <- data.frame(
  distance = c(170, 187, 179, 162, 187, 149, 171, 169, 161, 180, 167, 166, 172, 192,
              190, 210, 211, 202, 171, 176, 198, 187, 191, 175, 179, 177, 190, 179, 177,
              168, 196, 189, 187, 192, 200, 189, 201, 192, 180),
  nose = c("no", "yes", "yes", "yes", "no", "no", "no", "no", "yes", "yes", "no",
          "no", "no", "yes", "no", "yes", "yes", "no", "no", "yes", "yes", "no", "yes",
          "yes", "no", "no", "yes", "no", "no", "yes", "yes", "yes", "yes", "no", "yes",
          "yes", "yes", "yes", "no"),
  middle = c("no", "no", "no", "yes", "no", "yes", "yes", "yes", "no", "yes", "no",
             "no", "yes", "yes", "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "no",
             "no", "yes", "no", "no", "yes", "no", "yes", "no", "yes", "yes", "no", "yes",
             "yes", "yes", "no", "yes"),
  rear = c("yes", "yes", "yes", "yes", "no", "yes", "no", "yes", "no", "yes", "yes",
          "yes", "no", "no", "no", "no", "yes", "no", "yes", "no", "no", "yes", "no",
          "yes", "no", "yes", "yes", "no", "yes", "yes", "yes", "no", "no", "no", "yes",
          "no", "no", "yes"))
# Convert placement variables to factors
actual_data$nose <- factor(actual_data$nose)
actual_data$middle <- factor(actual_data$middle)
actual_data$rear <- factor(actual_data$rear)
```

## Linear Regression Analysis of Clip Placement Effects on Flight Distance

To analyze how paperclip placement affects the flight distance of paper airplanes, I conducted a linear regression using the lm() function. I chose to use linear regression with lm() rather than aov() because this study does not follow a Randomized Complete Block Design (RCBD) or Generalized Randomized Block Design (GRBD), where typically only one primary factor is of interest and interaction terms are often not included. In contrast, factorial designs like mine specifically focus on evaluating main effects and interactions, which lm() handles more flexibly. Since interaction terms are present and of interest in this study, lm() seemed like the appropriate modeling choice.

```
# Fit linear regression model with main effects and interactions
actual_model <- lm(distance ~ nose * middle * rear, data = actual_data)

# Extract summary coefficients
coef_summary <- summary(actual_model)$coefficients

# Round and convert to character
coef_table <- signif(coef_summary, 4)
coef_table <- as.data.frame(coef_table)

kable(coef_table,
      caption = "Regression Coefficients for Clip Placement Effects on Distance",
      digits = 4)
```

Table 4: Regression Coefficients for Clip Placement Effects on Distance

|                           | Estimate | Std. Error | t value | Pr(> t ) |
|---------------------------|----------|------------|---------|----------|
| (Intercept)               | 183.6    | 4.420      | 41.540  | 0.0000   |
| noseyes                   | 6.4      | 6.250      | 1.024   | 0.3135   |
| middleyes                 | -7.6     | 6.250      | -1.216  | 0.2329   |
| rearyes                   | -15.0    | 6.250      | -2.400  | 0.0224   |
| noseyes:middleyes         | 20.8     | 8.839      | 2.353   | 0.0249   |
| noseyes:rearyes           | 10.4     | 8.839      | 1.177   | 0.2480   |
| middleyes:rearyes         | 10.0     | 8.839      | 1.131   | 0.2663   |
| noseyes:middleyes:rearyes | -28.8    | 12.500     | -2.304  | 0.0279   |

Interpretations:

- The rear clip placement had a statistically significant effect on distance with a p-value of 0.0224, which is less than the significance level of 0.05. This provides sufficient evidence to reject the null hypothesis with respect to the rear clip, so the rear clip does impact performance. On average, placing a clip on the rear reduced flight distance by approximately 15 cm.
- The interaction between nose and middle placements was also statistically significant with a p-value of 0.0249, suggesting that the effect of placing a clip on the nose depends on whether there is also a clip in the middle. Since the p-value is below 0.05, we again reject the null hypothesis for this interaction, so this means the effect of placing a clip on the nose depends on whether there is also a clip in the middle
- The three-way interaction between nose, middle, and rear placements was also significant with a p-value of 0.0279, indicating a combined effect of all three clip positions. We reject the null hypothesis for this interaction.

These results highlight the importance of considering interaction effects in a factorial design. While individual placements may not significantly affect flight distance alone, certain combinations of placements do.

## Bonferroni Correction for Multiple Comparisons

Because the linear regression model includes multiple hypothesis tests for main effects and interactions, I applied a Bonferroni correction. The Bonferroni-adjusted alpha is  $.05/7 = .007$ . The Bonferroni adjustment is known to be conservative, so it is smaller than it needs to be in order to maintain the overall Type I Error rate. All p-values from the `lm()` output are greater than .007, so none of the predictors remain significant at the adjusted alpha. This means that after correcting for multiple testing, there is not enough statistical evidence to confidently claim any one predictor or interaction is significant.

## Checking Assumptions

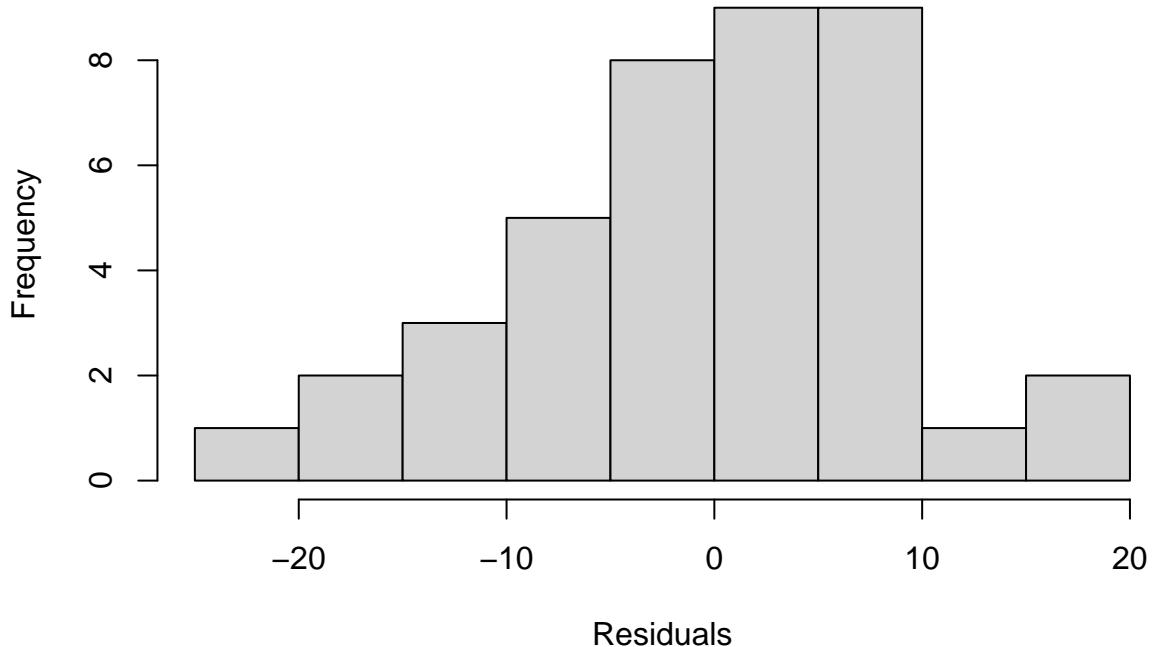
To ensure my interpretations of the results of the linear model are valid, it's important to check whether the underlying assumptions are reasonably satisfied.

### Normality of residuals

```
residuals_lm <- residuals(actual_model)

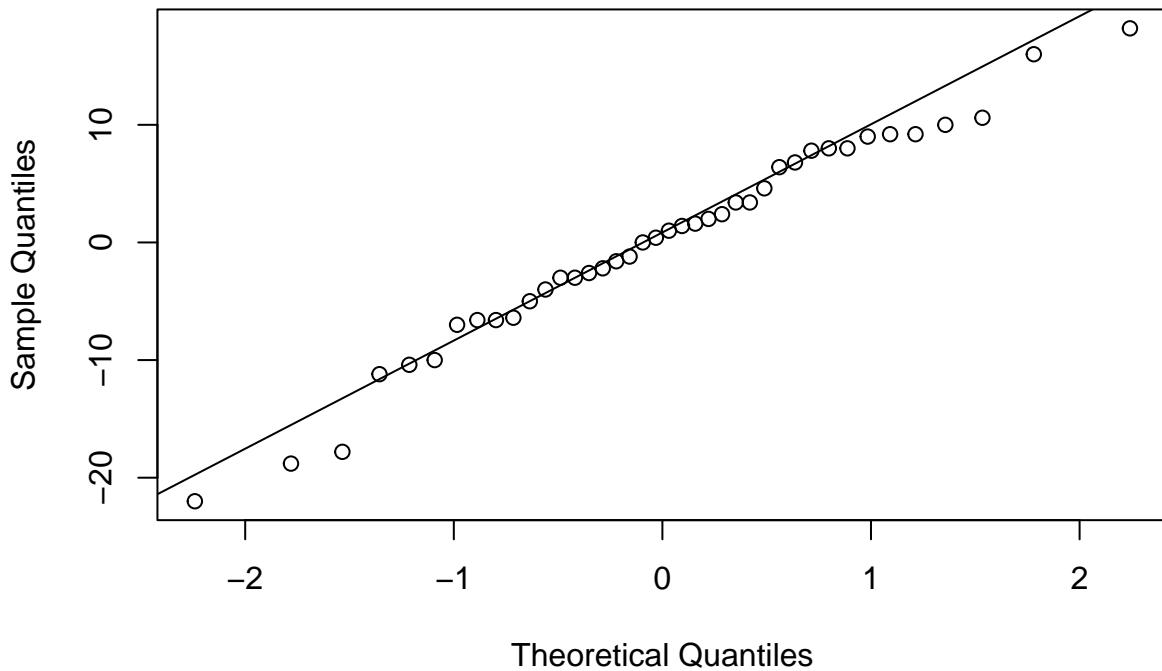
# Histogram of residuals
hist(residuals_lm, main = "Histogram of Residuals", xlab = "Residuals")
```

## Histogram of Residuals



```
# QQ plot for residuals
qqnorm(residuals_lm, main = "QQ Plot of Residuals")
qqline(residuals_lm)
```

## QQ Plot of Residuals



```
# Shapiro-Wilk test for normality
shapiro <- shapiro.test(residuals_lm)
shapiro_df <- data.frame(
  statistic = shapiro$statistic,
  pvalue = shapiro$p.value )
kable(shapiro_df, caption = "Shapiro-Wilk Test For Residuals")
```

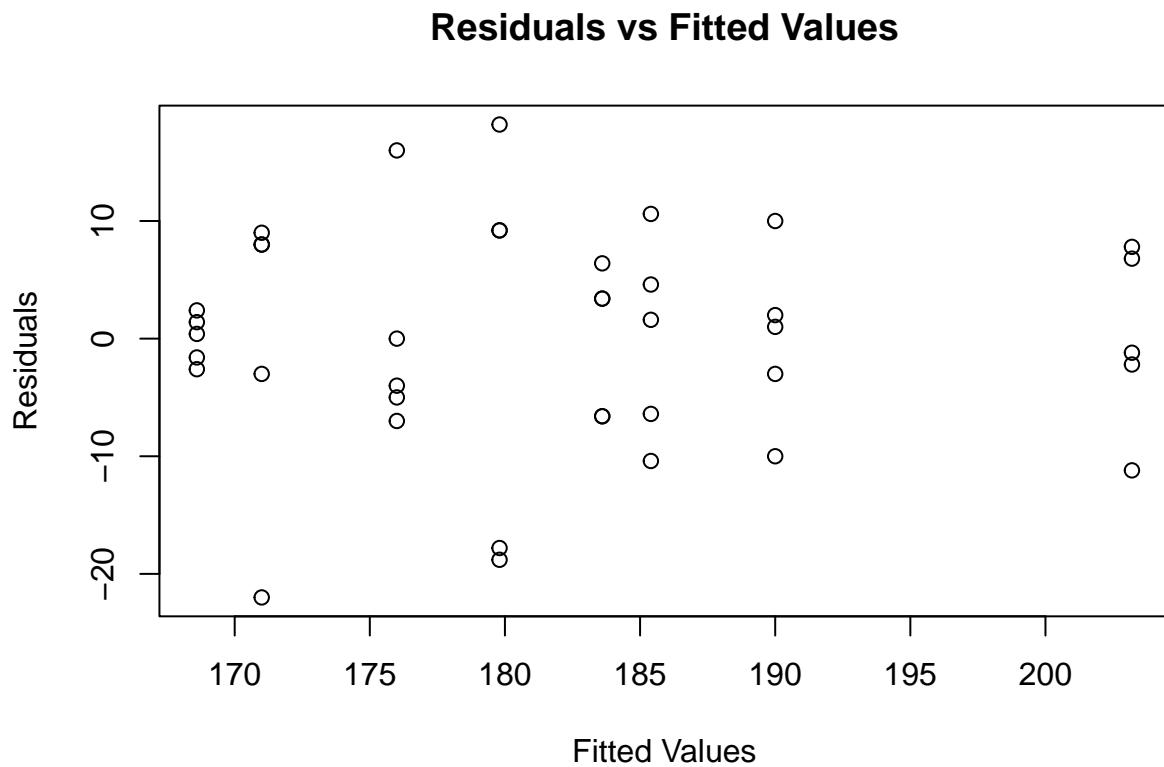
Table 5: Shapiro-Wilk Test For Residuals

|   | statistic | pvalue    |
|---|-----------|-----------|
| W | 0.9775951 | 0.6008822 |

The Shapiro-Wilk test for normality yielded a p-value of 0.6, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis that the residuals are normally distributed. This indicates that there is no strong evidence to suggest a significant deviation from normality. Visual inspections of the residuals further support this conclusion. The histogram of residuals appears approximately symmetric and does not show obvious signs of skewness. Additionally, the QQ-plot shows that most residual points closely follow the 45-degree reference line. However, some deviation is observed in the extreme upper and lower tails, where points stray from the line. This suggests the presence of potential outliers or heavier tails in the residual distribution, but they do not appear severe enough to violate the normality assumption.

## Equal Variances Across All Fitted Values

```
# Create a Residuals vs Fitted values plot
fitted_values <- fitted(actual_model)
plot(fitted_values, residuals_lm,
      xlab = "Fitted Values",
      ylab = "Residuals",
      main = "Residuals vs Fitted Values")
```

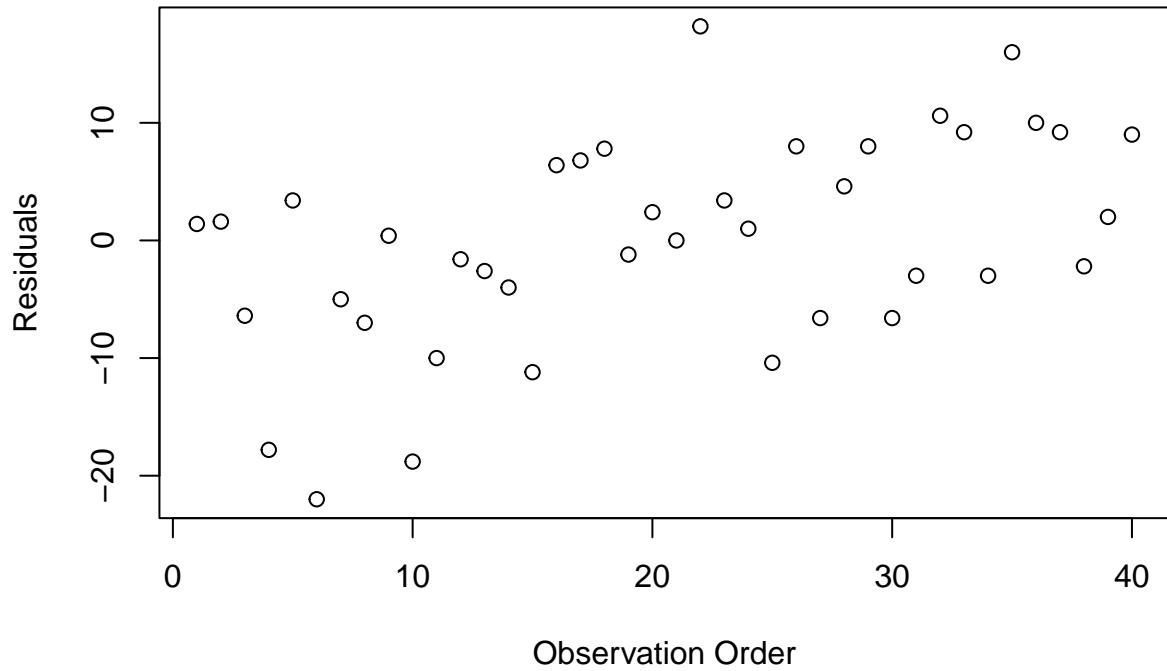


The plot of residuals versus fitted values shows that the residuals are relatively evenly dispersed around zero without any clear pattern or trend. This random scatter indicates that the variance of the residuals is approximately constant across different fitted values. Therefore, the assumption of equal variances of residuals appears to be satisfied in this model.

## Structure to the Data

```
# Residuals over order of data collection
time <- 1:length(residuals_lm)
plot(time, residuals_lm,
      xlab = "Observation Order",
      ylab = "Residuals",
      main = "Residuals vs. Observation Order")
```

## Residuals vs. Observation Order



The scatter plot of observations versus residuals reveals a random distribution of points with no discernible structure or pattern. This randomness indicates that the residuals are independent of the order of the throws, supporting the assumption that each throw was performed independently.

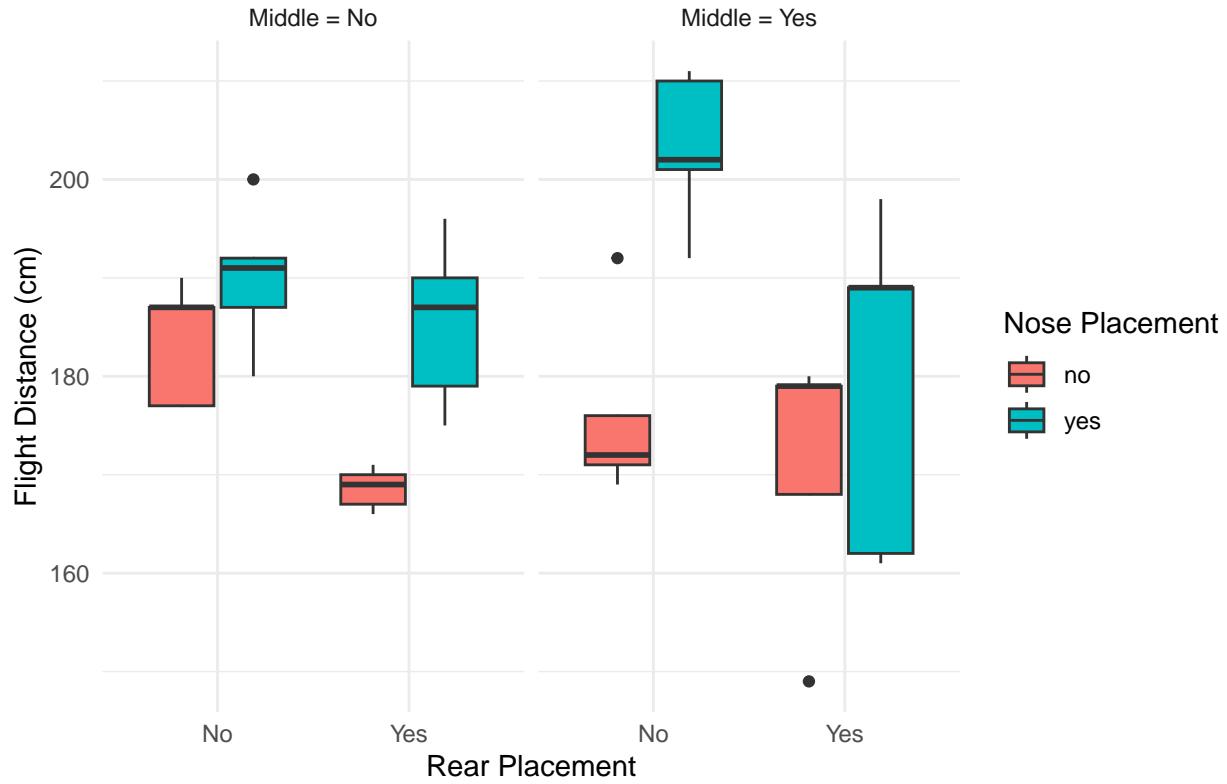
All assumptions of the linear model were met, so a permutation test was not necessary. Therefore, I relied on the p-values obtained from the lm() model.

## Visualizing the Results

### Boxplot of Flight Distance by Paperclip Placement

```
ggplot(actual_data, aes(x = rear, y = distance, fill = nose)) +
  geom_boxplot() +
  facet_wrap(~ middle,
             labeller = labeller(middle = c("yes" = "Middle = Yes", "no" = "Middle = No")) +
  scale_x_discrete(name = "Rear Placement", labels = c("yes" = "Yes", "no" = "No")) +
  labs(title = "Paper Airplane Distance Distributions by Paperclip Placement",
       x = "Rear Placement",
       y = "Flight Distance (cm)",
       fill = "Nose Placement") +
  theme_minimal()
```

## Paper Airplane Distance Distributions by Paperclip Placement

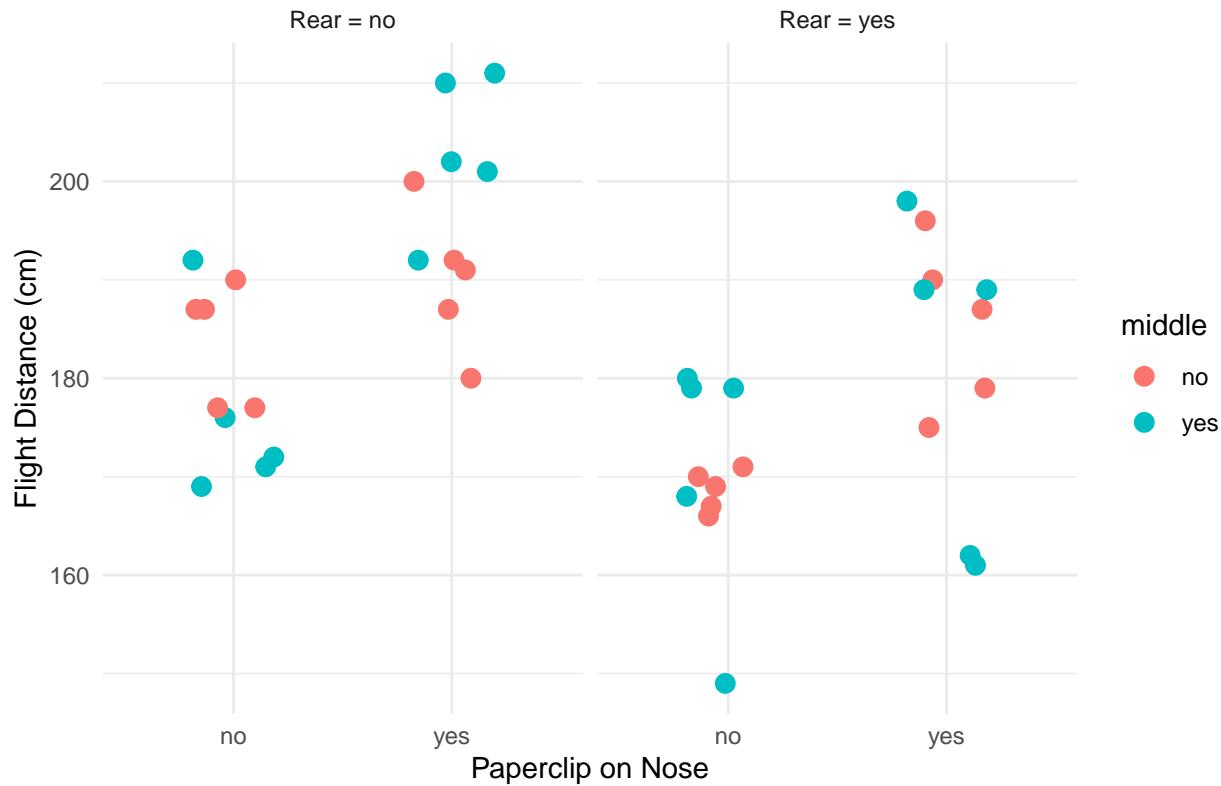


Here, we can see that when a clip is placed on both the middle and nose, flight distances tend to increase, with this combination showing the highest distances overall. The combination of clips on the nose, middle, and rear also exhibits the greatest variability in flight distances. In contrast, the scenario with no clip on the middle but a clip on the nose shows the least variability. Generally, clipping the rear results in shorter flight distances, while clipping the nose consistently leads to longer flight distances compared to conditions without the nose clip. These observed trends align with the linear regression analysis results.

### Scatter plot of Flight Distance by Paperclip Placement

```
ggplot(actual_data, aes(x = nose, y = distance, color = middle)) +
  geom_jitter(width = 0.2, height = 0, size = 3) +
  facet_grid(cols = vars(rear),
             labeller = labeller(rear = c("no" = "Rear = no", "yes" = "Rear = yes"))) +
  labs(
    title = "Scatter Plot of Flight Distance by Paperclip Placement",
    x = "Paperclip on Nose",
    y = "Flight Distance (cm)"
  ) +
  theme_minimal()
```

## Scatter Plot of Flight Distance by Paperclip Placement



The scatter plot shows that planes with a paperclip on the nose generally achieve longer flight distances. Additionally, planes without a paperclip on the rear tend to fly farther compared to those with rear placement. In contrast, the presence of a paperclip in the middle does not reveal any clear pattern in flight distance. Overall, the plot suggests that having a paperclip on the nose and none on the rear is associated with the longest flight distances.

## Discussion

This experiment evaluated how placing paperclips at the nose, middle, and rear of a paper airplane affects its flight distance, using a full  $2^3$  factorial design analyzed with a linear model. The model included all main effects and interaction terms to assess both individual and combined influences of clip placements. The results indicate that rear clip placement significantly reduced flight distance, while nose clip placement had a generally positive effect both visually in plots and statistically. Nose-only placements had some of the longest flights and lowest variability. Middle placement, on its own, did not yield a statistically significant change, and the scatter plot did not reveal a clear pattern. However, its role became more influential when combined with nose placement.

Notably, I found both a significant two-way interaction between nose and middle placement and a significant three-way interaction between all three clip positions. These interactions suggest that the influence of one clip location often depends on the presence or absence of others. Therefore, the effects of clip placement are more than just the sum of individual effects. When clips were placed simultaneously on the nose, middle, and rear, flight performance became much more variable, likely because the combined weight disrupted the airplane's balance. This confirms earlier research indicating that the way weight is distributed affects flight performance in combined ways that cannot be predicted by examining each weight placement separately.

It is clear that “adding paper clips changes the weight distribution and can improve stability, affecting lift and

drag during flight” (Learning Corner, n.d.). In real-world engineering, designers must consider these same effects when developing aircraft. Failing to properly manage weight distribution can compromise passenger safety and reduce flight efficiency due to increased fuel consumption or instability. Therefore, careful weight allocation is essential to achieve the most efficient aircraft performance. As Alexander (2001) explains, “Load your airplane improperly, and it will affect its fuel consumption, speed, rate of climb, controllability, ceiling, and even structural integrity.” Through this experiment, I gained valuable insight into how different weight placements influence flight times. I approached the problem like engineers tackling the aircraft weight and balance problem (WBP), which “involves the assignment of a set of unit load devices (ULDs) to a set of positions representing different holds in an aircraft under its weight and balance constraints” (Zhao et al., 2021).

The sample size of 5 replicates per treatment for a total of 40 throws was determined based on a pilot study and power analysis. Given that statistically significant results were obtained for multiple factors, this sample size was sufficient for the chosen design and variability level. However, after applying a Bonferroni correction, none of the individual predictors or interaction terms remained statistically significant. This outcome highlights the importance of correcting for multiple comparisons to reduce the risk of false positives, especially in models with many interaction terms. Given the limited sample size and the conservative nature of the Bonferroni correction, it is possible that the effects were not detected due to insufficient statistical power. A more appropriate next step would be to repeat the experiment with a larger sample size.

I could trust the p-values from the linear regression analysis and not need to perform a permutation test because the key statistical assumptions were satisfied. Normality of residuals was supported both visually and through a Shapiro-Wilk test, indicating no significant departure from normality. Equal variances across all fitted values appeared to be satisfied because the residuals vs. fitted values plot showed no clear pattern. Lastly, independence of observations was supported by a residuals vs. order of data collection plot, which showed a random scatter.

While these assumptions validate the use of linear regression, several limitations of the experiment must be acknowledged. Despite maintaining a consistent technique, human throwing introduces natural variability. Also, although I conducted this experiment indoors, minor air disturbances may have influenced results, and repeated use of a single plane improved consistency but could have introduced degradation over time. Clearly, generalization is limited to the specific conditions used here: a particular airplane design, paper type, clip size, and throwing method.

Results may vary under different conditions, so future studies should investigate a range of plane designs, clip types, material stiffness, and environmental factors. For example, this study used clips of only one weight; using both lighter and heavier clips in future experiments could provide more insight into how mass distribution influences flight. Additionally, using automated launch mechanisms would prevent variability from throwing. Lastly, detailed tracking of flight paths using video analysis would allow me to analyze the observed effects more thoroughly. I could link specific clip placements to changes in flight trajectories, rather than relying solely on flight distance.

## References

- Alexander, R. (2001, March 1). Weight & balance: Weighing and measuring your safety. Experimental Aircraft Association. <https://www.eaa.org/eaa/aircraft-building/builderresources/while-youre-building/building-articles/weight-and-balance/weight-and-balance>
- Kundu, A. K. (2010). Aircraft Design. Cambridge: Cambridge University Press.
- Learning Corner. (n.d.). Understanding the effects of paper clips on paper plane performance. Retrieved June 5, 2025, from <https://learningcorner.co/explain-anything/6869>
- Raymer, D. (2012). Aircraft design: A conceptual approach (5th ed.). American Institute of Aeronautics and Astronautics. ISBN: 978-1-60086-911-2.

Torenbeek, E. (1982). Airplane weight and balance. In: Synthesis of Subsonic Airplane Design. Springer, Dordrecht. [https://doi.org/10.1007/978-94-017-3202-4\\_8](https://doi.org/10.1007/978-94-017-3202-4_8)

Zhao, X., Yuan, Y., Dong, Y., & Zhao, R. (2021). Optimization approach to the aircraft weight and balance problem with the centre of gravity envelope constraints. IET Intelligent Transport Systems. <https://doi.org/10.1049/itr2.12096>