How an ANOVA IS A REGRESSION WIHT DUMMY VARIABLES

Bill Perry

# 1. Introduction

This document demonstrates how Analysis of Variance (ANOVA) is mathematically equivalent to a regression model with dummy variables using an example with R code and visualizations.

# 2. Setup and Data Creation

Let’s begin by loading necessary packages and creating a dataframe about plant heights with three different fertilizer treatments.

# install.packages("flextable")  
library(tidyverse)  
library(flextable)  
  
# Create the dataset  
fertilizer\_data <- tibble(  
 fertilizer = rep(c("A", "B", "C"), each = 3),  
 height = c(10, 12, 8, # Fertilizer A  
 14, 16, 18, # Fertilizer B  
 20, 22, 24) # Fertilizer C  
)  
  
# Display the dataset using flextable  
flextable(fertilizer\_data) %>%  
 set\_caption("Plant Heights by Fertilizer Type") %>%  
 theme\_vanilla() %>%  
 autofit()

| **fertilizer** | **height** |
| --- | --- |
| A | 10 |
| A | 12 |
| A | 8 |
| B | 14 |
| B | 16 |
| B | 18 |
| C | 20 |
| C | 22 |
| C | 24 |

# 3. Calculating Group Means (ANOVA Approach)

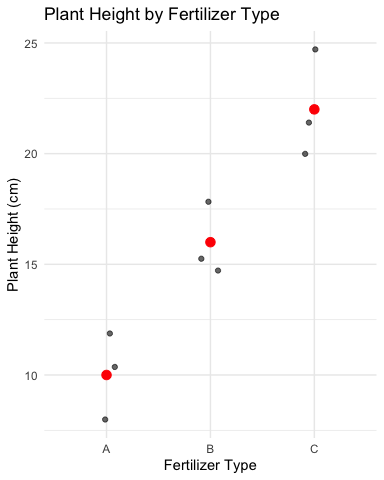
In ANOVA, we calculate the mean of each group and compare variation between groups to variation within groups.

group\_means <- fertilizer\_data %>%  
 group\_by(fertilizer) %>%  
 summarize(mean\_height = mean(height))  
  
flextable(group\_means) %>%  
 set\_caption("Group Means (ANOVA Approach)") %>%  
 theme\_vanilla() %>%  
 autofit()

| **fertilizer** | **mean\_height** |
| --- | --- |
| A | 10 |
| B | 16 |
| C | 22 |

Let’s visualize the raw data and group means:

ggplot(fertilizer\_data, aes(x = fertilizer, y = height)) +  
 geom\_jitter(width = 0.1, alpha = 0.6) +  
 geom\_point(data = group\_means, aes(y = mean\_height),   
 color = "red", size = 3) +  
 labs(title = "Plant Height by Fertilizer Type",  
 x = "Fertilizer Type",  
 y = "Plant Height (cm)") +  
 theme\_minimal()



# 4. Running the ANOVA

# Run ANOVA  
anova\_model <- aov(height ~ fertilizer, data = fertilizer\_data)  
anova\_summary <- summary(anova\_model)  
anova\_summary

Df Sum Sq Mean Sq F value Pr(>F)   
fertilizer 2 216 108 27 0.001 \*\*  
Residuals 6 24 4   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# 5. Regression with Dummy Variables

For the regression approach, we’ll create dummy variables for fertilizer types, using fertilizer A as the reference level.

# Set fertilizer A as the reference level  
fertilizer\_data$fertilizer <- factor(fertilizer\_data$fertilizer, levels = c("A", "B", "C"))  
  
# Run regression with dummy variables  
reg\_model <- lm(height ~ fertilizer, data = fertilizer\_data)  
reg\_summary <- summary(reg\_model)  
reg\_summary

Call:  
lm(formula = height ~ fertilizer, data = fertilizer\_data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -2 -2 0 2 2   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 10.000 1.155 8.660 0.000131 \*\*\*  
fertilizerB 6.000 1.633 3.674 0.010402 \*   
fertilizerC 12.000 1.633 7.348 0.000325 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2 on 6 degrees of freedom  
Multiple R-squared: 0.9, Adjusted R-squared: 0.8667   
F-statistic: 27 on 2 and 6 DF, p-value: 0.001

# 6. Understanding the Regression Coefficients

In our regression model:

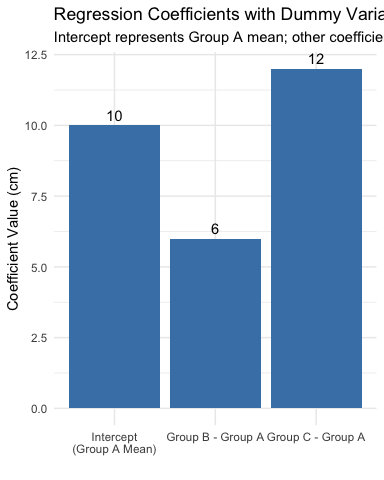
* The intercept (10) is equal to the mean of the reference group (A)
* The coefficient for fertilizer B (6) is the difference between mean of group B and mean of group A
* The coefficient for fertilizer C (12) is the difference between mean of group C and mean of group A

# Create a table showing the relationship between coefficients and means  
coefs <- coef(reg\_model)  
  
coefficients\_explained <- tibble(  
 Term = c("Intercept", "fertilizerB", "fertilizerC"),  
 Coefficient = coefs,  
 Meaning = c(  
 "Mean of Group A (reference group)",  
 "Difference between Group B and Group A means",  
 "Difference between Group C and Group A means"  
 ),  
 Mathematical\_Expression = c(  
 "β₀ = μA",  
 "β₁ = μB - μA",  
 "β₂ = μC - μA"  
 ),  
 Numeric\_Value = c(coefs[1],  
 paste0(round(group\_means$mean\_height[2], 1), " - ",   
 round(group\_means$mean\_height[1], 1), " = ",   
 round(coefs[2], 1)),  
 paste0(round(group\_means$mean\_height[3], 1), " - ",   
 round(group\_means$mean\_height[1], 1), " = ",   
 round(coefs[3], 1))))  
  
# Use flextable to format the table  
flextable(coefficients\_explained) %>%  
 set\_caption("Regression Coefficients Explained") %>%  
 theme\_vanilla() %>%  
 fit\_to\_width(max\_width = 8, unit = "in") %>%  
 bold(j = 1) %>%  
 colformat\_double(j = 2, digits = 2)

| **Term** | **Coefficient** | **Meaning** | **Mathematical\_Expression** | **Numeric\_Value** |
| --- | --- | --- | --- | --- |
| **Intercept** | 10.00 | Mean of Group A (reference group) | β₀ = μA | 10 |
| **fertilizerB** | 6.00 | Difference between Group B and Group A means | β₁ = μB - μA | 16 - 10 = 6 |
| **fertilizerC** | 12.00 | Difference between Group C and Group A means | β₂ = μC - μA | 22 - 10 = 12 |

Let’s visualize these coefficients:

coef\_data <- tibble(  
 Term = factor(c("Intercept\n(Group A Mean)", "Group B - Group A", "Group C - Group A"),  
 levels = c("Intercept\n(Group A Mean)", "Group B - Group A", "Group C - Group A")),  
 Value = c(coefs[1], coefs[2], coefs[3])  
)  
  
ggplot(coef\_data, aes(x = Term, y = Value)) +  
 geom\_col(fill = "steelblue") +  
 geom\_text(aes(label = round(Value, 1)), vjust = -0.5) +  
 labs(title = "Regression Coefficients with Dummy Variables",  
 subtitle = "Intercept represents Group A mean; other coefficients show differences from reference",  
 x = "",  
 y = "Coefficient Value (cm)") +  
 theme\_minimal()



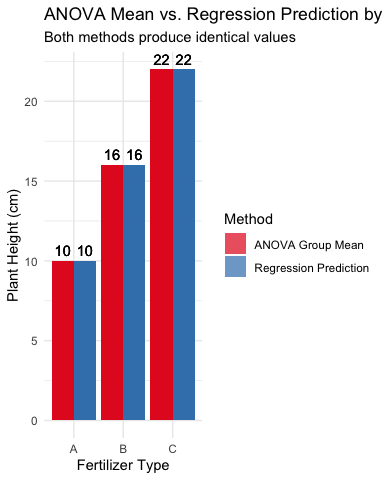
# 7. Demonstrating the Equivalence

Now, let’s prove that the regression model predictions are identical to the ANOVA group means:

# Get predictions from regression model  
predicted\_values <- predict(reg\_model, fertilizer\_data)  
  
# Create a dataframe for comparison  
comparison\_data <- fertilizer\_data %>%  
 mutate(predicted = predicted\_values) %>%  
 group\_by(fertilizer) %>%  
 mutate(group\_mean = mean(height))  
  
# Generate the predicted values for each group  
predicted\_values\_by\_group <- comparison\_data %>%  
 group\_by(fertilizer) %>%  
 reframe(  
 anova\_mean = mean(height),  
 regression\_prediction = mean(predicted),  
 formula = case\_when(  
 fertilizer == "A" ~ paste0(round(coefs[1], 1), " + 0 + 0 = ", round(coefs[1], 1)),  
 fertilizer == "B" ~ paste0(round(coefs[1], 1), " + ", round(coefs[2], 1), " + 0 = ", round(coefs[1] + coefs[2], 1)),  
 fertilizer == "C" ~ paste0(round(coefs[1], 1), " + 0 + ", round(coefs[3], 1), " = ", round(coefs[1] + coefs[3], 1))  
 )  
 )

Let’s visualize this equivalence:

# Create data for plotting the equivalence  
plot\_data <- predicted\_values\_by\_group %>%  
 pivot\_longer(cols = c(anova\_mean, regression\_prediction),  
 names\_to = "method",  
 values\_to = "value") %>%  
 mutate(method = ifelse(method == "anova\_mean", "ANOVA Group Mean", "Regression Prediction"))  
  
ggplot(plot\_data, aes(x = fertilizer, y = value, fill = method)) +  
 geom\_bar(stat = "identity", position = position\_dodge(), alpha = 0.7) +  
 geom\_text(aes(label = round(value, 1)), position = position\_dodge(width = 0.9), vjust = -0.5) +  
 labs(title = "ANOVA Mean vs. Regression Prediction by Fertilizer Type",  
 subtitle = "Both methods produce identical values",  
 x = "Fertilizer Type",  
 y = "Plant Height (cm)",  
 fill = "Method") +  
 theme\_minimal() +  
 scale\_fill\_brewer(palette = "Set1")



# 8. Comparing Statistical Tests

Both ANOVA and regression provide an F-test. Let’s compare them:

# ANOVA: Extract F-value and p-value  
anova\_f <- anova\_summary[[1]]$`F value`[1]  
anova\_p <- anova\_summary[[1]]$`Pr(>F)`[1]  
  
# Regression: Extract F-value and p-value  
reg\_f <- reg\_summary$fstatistic[1]  
reg\_p <- pf(reg\_f, reg\_summary$fstatistic[2], reg\_summary$fstatistic[3], lower.tail = FALSE)  
  
# Compare them  
test\_comparison <- tibble(  
 Test = c("ANOVA F-test", "Regression F-test"),  
 `F-value` = c(anova\_f, reg\_f),  
 `p-value` = c(anova\_p, reg\_p)  
)  
  
# Format with flextable  
flextable(test\_comparison) %>%  
 set\_caption("Comparison of Statistical Tests") %>%  
 theme\_vanilla() %>%  
 autofit() %>%  
 colformat\_double(j = 2:3, digits = 4)

| **Test** | **F-value** | **p-value** |
| --- | --- | --- |
| ANOVA F-test | 27.0000 | 0.0010 |
| Regression F-test | 27.0000 | 0.0010 |

# 9. The Mathematical Relationship

For a one-way ANOVA with a categorical variable having k levels, we can express the relationship with regression as:

Where: - is the mean of the reference group - are the differences between each group’s mean and the reference group mean - are dummy variables (0 or 1)

In our example: - (mean of group A) - (difference between B and A) - (difference between C and A)

# 10. Conclusion

This demonstration shows that one-way ANOVA is mathematically equivalent to regression with dummy variables. The key equivalences are:

1. ANOVA group means = Regression predictions for each group
2. F-statistic from ANOVA = F-statistic from regression
3. p-values are identical in both approaches

This confirms that both techniques are special cases of the General Linear Model, just expressed in different ways. For a categorical predictor with k levels, we need k-1 dummy variables in the regression approach, with one level serving as the reference category.