Common Statistical Models

VT-DSPG 2025

Anbin Rhee

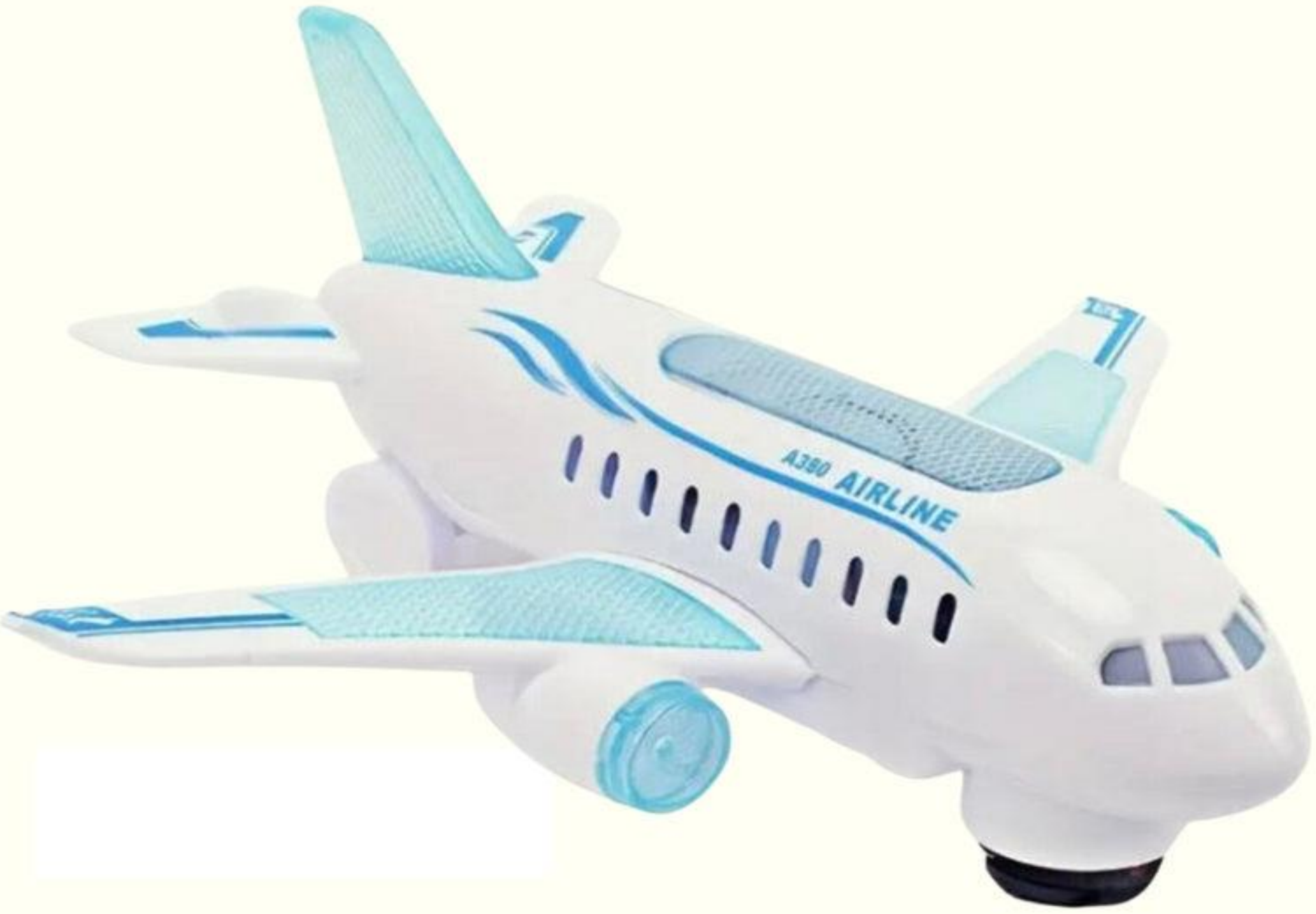
July 2, 2025

Contents

1. What is a Model?
2. Simple Linear Regression
3. Multiple Linear Regression
4. ANOVA
5. Logistic Regression

What is a “Model”?

- Analogy: A toy airplane
  - A toy airplane is not a real plane, but it helps us understand how a real plane flies
- Statistical Models are similar
  - They are simplified mathematical representations of complex real-world data
  - They help us understand relationships between variables and make predictions



Key Concepts: Variables

- What is a variable?
  - Anything that can be measured or observed and can vary across observations
- Independent Variable (IV) : Predictor / Cause
  - The variable we change or observe to see if it has an effect

    (The amount of fertilizer you give a plant)
- Dependent Variable (DV) : Outcome / Effect
  - The variable we measure to see if it changes in response to the independent variable

    (The height of the plant)

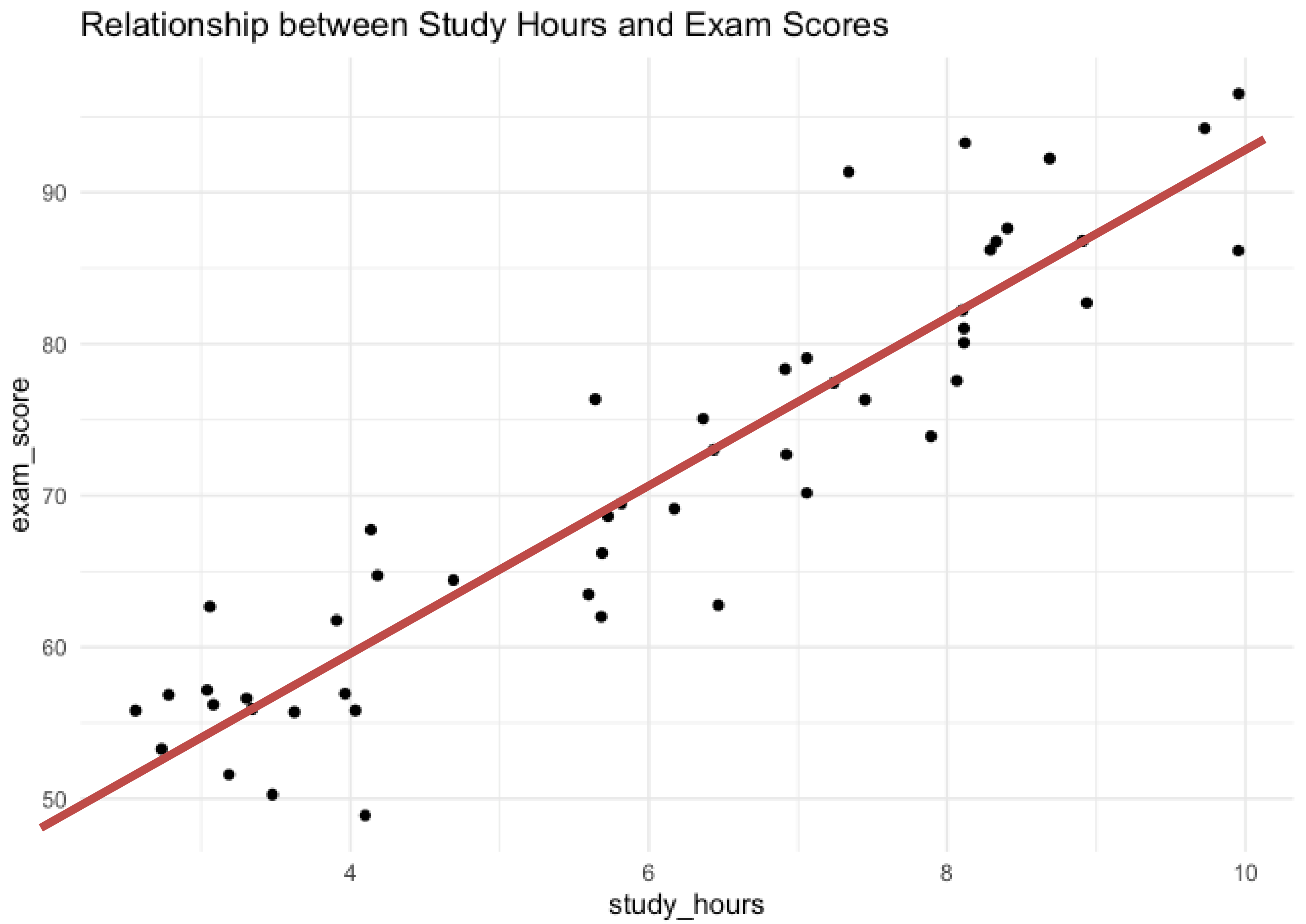
Key Concepts: Variable Types

- Numerical
  - Can be any value within a range
    - Continuous : can take any values withing a range
    - Discrete : can only take specific, separate values
  - e.g., temperature, sales, study hours, the number of children
- Categorical
  - Falls into distinct categories
  - e.g., gender, product type, teaching method (A/B/C)

Simple Linear Regression (SLR)

Example: Can we predict a student's **Exam Score** based on their **Study Hours**?

- We want to predict **a continuous numerical outcome** using **one continuous numerical predictor**



Simple Linear Regression

- The Conceptual Model (The 'True' Relationship in the population) :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

  - Y_i : The true outcome (Dependent variable) for individual i (e.g., Exam Score)
  - X_i : The predictor (Independent variable) for individual i (e.g., Study Hours)
  - β_0 : The true average value of Y when X is 0 (Intercept)
  - β_1 : The true average change in Y for every one-unit increase in X (Slope)
  - ϵ_i : The part of Y_i that the model cannot explain for individual i (Error term)

Simple Linear Regression

- The Estimated Model (From Our Data) :

$$\hat{Y}_i = \widehat{\beta_0} + \widehat{\beta_1}\, X_i$$

  - $\hat{Y}_i$ (Y-hat) : The predicted outcome for individual i (e.g., Exam Score)
  - $\widehat{\beta_0}$ (Estimated Intercept) : Our best guess for β_0 based on our data
  - $\widehat{\beta_1}$ (Estimated Slope) : Our best guess for β_1 based on our data

- Error vs. Residual
  - Error (ϵ_i): Unobserved difference between the true Y_i and the true model's prediction
  - Residual (e_i) : Observed difference between the actual Y_i and our model's predicted $\hat{Y}_i$

$e_i = Y_i - \hat{Y}_i$: We want small residuals!

Simple Linear Regression: Assumptions

1. Independence of Errors
   - Errors (ϵ_i) are independent of each other
2. Linearity
   - The relationship between X and Y is approximately linear
3. Constant Variance of Errors (Homoscedasticity)
   - The variability of the residuals should be roughly constant across all levels of the predictor X
4. Normality
   - The residuals (e_i) are approximately normally distributed

Simple Linear Regression: Standard Error

- Standard Error (SE) of Coefficients:
  - How much our estimates would likely vary if we took many different samples from the same population
  - A smaller SE means our estimate is more precise and reliable (less-sample to sample variation)

Simple Linear Regression: T-Test

- Hypotheses

$$H_0: \beta_1 = 0 \text{ vs. } H_1: \beta_1 \neq 0$$

- We need statistical test used to get the p-value for individual coefficients in regression
- It calculates a t-statistics for each coefficient:

$$t = \frac{Coefficient\ Estimate\ (\beta_1)\ - Hypothesized\ Value\ (\beta_1 = 0\ under\ H_0)}{Standard\ Error\ of\ Coefficient\ (\beta_1)}$$

  - A smaller SE means our estimate is more precise and reliable
- It measures how many standard errors away our estimated coefficient is from the hypothesized value
- A **large** |t| means our estimate is further from zero, making it **less** likely to be due to chance

SLR: Confidence Interval (CI)

- Instead of just a single point estimate, a CI provides a range of plausible values for the true population parameters
- Example: A 95% CI for the slope might be (a, b). This means, “We are 95% confident that the true average change in Y for a one-unit change in X lies somewhere between a and b”
- It gives us a sense of the precision of our estimate and the range of values we’d expect the true parameters to fall in if we repeated the experiment many times

Simple Linear Regression: Interpretation

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.6210     2.1316   17.65   <2e-16 ***
study_hours   5.5231     0.3314   16.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.13 on 48 degrees of freedom
Multiple R-squared:  0.8526,    Adjusted R-squared:  0.8495 
F-statistic: 277.7 on 1 and 48 DF,  p-value: < 2.2e-16
```

- Intercept(37.6210): The predicted exam score for a student who studied 0 hours
- Study_hours (5.5231): For every additional hour of study, the exam score increases on average by 5.52 points

Multiple Linear Regression

Example: Can we predict an individual's Salary based on their Year of Experience, Education Level, and Hours worked Per Week?

• Dependent variable : Salary		*Continuous*
• Independent variables :	Year of Experience,	*Continuous*
	Education Level	*Categorical*
	Hours Worked per Week	*Continuous*

- We want to predict/explain a **continuous numerical outcome** using two or more independent variables
- Extends simple linear regression by adding more predictors

Multiple Linear Regression

- The Conceptual Model (The 'True' Relationship in the population) :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

  - It is adding more X terms to our simple linear regression equation
  - Y_i : Salary for individual i
  - X_{1i} : Years of Experience for individual i
  - X_{2i} : Education Level for individual i (High school, Bachelor, Master, PhD)
  - X_{3i} : Hours Worked per Week for individual i
- Each β_j (slope) now represents the average change in Y for a one-unit change in that specific X, while holding all other X variables constant.

Handling Categorical Predictors

- For handling categorical predictors, one category becomes the reference group
- Statistical software automatically convert categorical variables into ‘**dummy variables**’ (0s and 1s) to indicate each category
- The coefficients for other categories show the predicted difference compared to this reference group
- For the salary example, we can set High School as our reference group for Education Level

Multiple Linear Regression: Interpretation

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)             35480.39    3987.61   8.898 4.05e-14 ***
experience               1907.66      85.39  22.341  < 2e-16 ***
education_levelBachelor  5375.48    1441.44   3.729 0.000328 ***
education_levelMaster   10936.71    1410.09   7.756 1.03e-11 ***
education_levelPhD      15381.78    1411.88  10.895  < 2e-16 ***
hours_per_week            115.86      93.38   1.241 0.217811    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4658 on 94 degrees of freedom
Multiple R-squared:  0.8648,	Adjusted R-squared:  0.8576 
F-statistic: 120.3 on 5 and 94 DF,  p-value: < 2.2e-16
```

- Experience (1907.66): Each additional year of experience increases predicted salary by about $1907, adjusting for education level and hours worked (very strong effect)
- Education level – Bachelor: $5375 more than High School, on average

Comparing Multiple Groups: ANOVA

Example: The company wants to compare the average purchase amounts among customers exposed to three different marketing campaigns (A, B, and C) to see which campaign performs best.

- Comparing the means of a **continuous Dependent Variable (purchase amount)** among **three marketing groups** (**Campaign A, B, C)**
- Analysis Of Variance (ANOVA) helps us:
  - Determine if there is a statistically significant difference among the group means

Comparing Multiple Groups: ANOVA

- The Conceptual Model:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

  - Y_{ij} : The outcome (Dependent variable) for individual i in j group (Purchase Amount)
  - μ : The overall average of Y across all groups (Grand Mean)
  - α_j : The effect of being in group j (Campaign A, B, or C). It represents how much the mean of group j deviates from the grand mean
  - ϵ_{ij} : The part of Y_{ij} that the model cannot explain

Comparing Multiple Groups: ANOVA

- Hypothesis:

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 \ \text{ vs. } H_1\text{: at least one group differ}$$

- ANOVA looks at two types of variation in the data:

  1. Between-Group Variation: The variability among the group means themselves (effect)
  2. Within-Group Variation: The variability within each group (noise)

- F-statistics:

$$F = \frac{Variability\ \ between\ \ Group\ \ Means}{Variability\ \ within\ \ Groups}$$

  - A larger ‘F’ value suggest the group differences are large compared to the random differences withing each group, indicating significant group effects

ANOVA: Interpretation

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   49.064      1.310  37.455  < 2e-16 ***
campaignB     13.121      1.853   7.083 1.03e-11 ***
campaignC      6.383      1.853   3.446 0.000652 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.1 on 297 degrees of freedom
Multiple R-squared:  0.1445,	Adjusted R-squared:  0.1388 
F-statistic: 25.09 on 2 and 297 DF,  p-value: 8.556e-11
```

- CampaignB (13.121): Customers in Campaign B had an average purchase amount that was $13.121 higher than customers in Campaign A. This difference is highly statistically significant

Logistic Regression

Example: The medical company has developed a new treatment to improve treatment success compared to the standard treatment, after accounting for patient age and gender

- Outcome (Y) is either success or failure *Binary*
- Predictors (X): Treatment group (New vs. Standard) *Categorical*

  Age *Continuous*

  Gender *Categorical*
- Linear regression predicts continuous values. It could predict "success" as -0.5 or 1.2, which don't make sense for a Yes/No outcome

Logistic Regression

- Instead of predicting the outcome (0 or 1) directly, Logistic Regression predicts the **probability of the 'Yes/Success' outcome** occurring (e.g., the probability of treatment success)
- The Model:

$$log(Odds\ of\ success) = \beta_0 + \beta_1 * Treatment + \beta_2 * Age + \beta_3 * Gender$$

- $Odds = \frac{Probability\ of\ Event}{1-Probability\ of\ Event} = \frac{Probability\ of\ Success}{Probability\ of\ Failure}$
- Purpose: To predict the probability of a binary outcome and understand which factors increase or decrease the odds of that outcome

Logistic Regression

- The Model:

$$log(Odds\ of\ success) = \beta_0 + \beta_1 * Treatment + \beta_2 * Age + \beta_3 * Gender$$

  - $Odds = \frac{Probability\ of\ Event}{1 - Probability\ of\ Event}$
- β_1 : change in log-odds of success for new treatment compared to standard
- Exp(β_1): **odds ratio of success** for new treatment
  - How many times higher or lower the odds of success are under the new treatment compared to the standard treatment
- Logistic regression models the **probability of success**, allowing adjustment for other factors (age, gender) while comparing treatment effects

Logistic Regression: Interpretation

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.01130    0.95185   1.062   0.2880
treatmentStandard -0.74231    0.35994  -2.062   0.0392 *
age               -0.03402    0.01900  -1.790   0.0735 .
genderMale        -0.55005    0.35980  -1.529   0.1263
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- TreatmentStandard (Estimate: -0.74231): Patients in the Standard treatment group have exp(-0.74231)=0.476 times the odds (or are about 52.4% less likely) of achieving treatment success compared to patients in the New treatment group, holding age and gender constant. This difference is statistically significant.

Quick Quiz

- What statistical model would you use to compare the average productivity across these three training programs?

Example: A company tests three types of training programs to improve employee productivity. After the training , they measure each employee’s number of tasks completed per day.

- **ANOVA model** to compare the mean productivity across three training program

Important Considerations

- Garbage in, garbage out
  - Bad data -> bad insight
- Correlation is NOT causation
  - Just because two variables are related does not mean one causes the other
- Models are Simplifications
  - They are mathematical abstractions of reality. They do not capture every nuance
- Overfitting
  - A model that is too complex for your data might fit your current data perfectly but perform poorly on new, unseen data

Summary

- **Simple Linear Regression**: Predicting a continuous outcome from one continuous predictor (e.g., Exam Score from Study Hours)
- **Multiple Linear Regression**: Predicting a continuous outcome from multiple predictors (continuous or categorical) (e.g., Salary from Experience, Education, Hours)
- **ANOVA**: Comparing the means of a continuous outcome across three or more groups (e.g., Purchase Amount from Campaign A, B, and C)
- **Logistic Regression**: Predicting a binary (Yes/No) outcome from one or more predictors (e.g., Treatment Success from Treatment Group, Age, Gender)