# STAT 333: Two-Sample $t$-Test as a Special Case of Linear Regression

Lecture Notes (Weeks 1–2)

February 18, 2025

## Overview

These notes combine:

- A student's handwritten notes from the first lecture on simple linear regression and its connection to the two-sample $t$-test.
- Blackboard work shown in lecture (images provided).
- A transcription of the instructor's second lecture, expanded and organized into coherent explanations.

We focus on how the two-sample $t$-test is a special case of simple linear regression and introduce why regression (including multiple regression) is important for understanding relationships (and differences) in data.

---

## 1 Lecture 1 Notes (Student Transcription)

### 1.1 Simple Linear Model

- We collect $(x_i, y_i)$ for $i = 1, \ldots, n$, where $x_i$ is a feature (or predictor) and $y_i$ is an outcome (or response).
- The **simple linear model** posits

$$y_i \;=\; \beta_0 + \beta_1 x_i \;+\; \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

- Our goal is to estimate the parameters $\beta_0$ (intercept) and $\beta_1$ (slope).

### 1.2 Recall: $t$-test

- In a one-sample $t$-test setting, we assume

$$y_i \;\sim\; \mathcal{N}(\mu, \sigma^2) \quad \text{i.i.d.},$$

and the null hypothesis is often $H_0 : \mu = 0$ (or some other specified value). The usual estimator of $\mu$ is

$$\hat{\mu} \;=\; \bar{y} \;=\; \frac{1}{n} \sum_{i=1}^{n} y_i.$$

### 1.3 Linear Model as a Generalization of the $t$-test

- **Two-sample $t$-test:** Suppose we have two groups (control vs. treatment). The two-sample $t$-test model can be written as:

$$y_i \sim \mathcal{N}(\mu_C, \sigma^2) \quad \text{(control group)}, \quad y_i \sim \mathcal{N}(\mu_T, \sigma^2) \quad \text{(treatment group)}.$$

  The null hypothesis is $H_0 : \mu_C = \mu_T$.

- To see this as a regression, define a *dummy variable*

$$x_i = \begin{cases} 0, & \text{if } i \text{ is in control,} \\ 1, & \text{if } i \text{ is in treatment.} \end{cases}$$

  Then the regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

  becomes

$$y_i = \beta_0 + \varepsilon_i \quad (\text{if } x_i = 0),$$
$$y_i = \beta_0 + \beta_1 + \varepsilon_i \quad (\text{if } x_i = 1),$$

  which matches the two-sample setup:

$$y_i \sim \mathcal{N}(\beta_0, \sigma^2) \quad \text{(group 0)}, \qquad y_i \sim \mathcal{N}(\beta_0 + \beta_1, \sigma^2) \quad \text{(group 1)}.$$

  Here, $\beta_0$ corresponds to $\mu_C$ and $\beta_0 + \beta_1$ corresponds to $\mu_T$.

### 1.4 Summary of Lecture 1

- Key insight: **Two-sample $t$-test** is indeed a **special case** of **simple linear regression** when $x_i \in \{0, 1\}$.
- Testing $H_0 : \beta_1 = 0$ in the regression is equivalent to testing $H_0 : \mu_C = \mu_T$ in the two-sample $t$-test.

---

## 2 Blackboard Highlights (End of Lecture 1 & Beginning of Lecture 2)

From the provided images, the board contained:

- **Threads from previous lecture:**

  "Two-sample $t$-test is a special case of linear regression."

- **Defining a 0–1 (dummy) variable for two groups:**

$$x_i = \begin{cases} 0 & \text{(control group)}, \\ 1 & \text{(treatment group)}. \end{cases}$$

- **Simple Linear Regression (SLR) model:**

$$Y_i \;=\; \beta_0 \;+\; \beta_1\, x_i \;+\; \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)\,\text{i.i.d.}$$

- **Equivalence to T-test distributions:**

$$\begin{aligned} Y_i &\sim \mathcal{N}(\beta_0,\, \sigma^2), && \text{if } x_i = 0 \text{ (control)},\\ Y_i &\sim \mathcal{N}(\beta_0 + \beta_1,\, \sigma^2), && \text{if } x_i = 1 \text{ (treated)}. \end{aligned}$$

- **Testing $H_0 : \beta_1 = 0$:**

  This is the same as asking if there's no difference between treatment and control means.

---

# 3 Lecture 2 (Instructor's Transcribed Discussion, Organized)

Below is a condensed and edited version of the instructor's spoken remarks, showing how they connect to the blackboard material above and expanding on the reasoning behind treating the $t$-test as a regression.

## 3.1 Recap: Two-Sample $t$-Test and Simple Linear Regression

- At the start of Lecture 2, the instructor asked the class to recall "threads" from the previous lecture. The main takeaway: *the two-sample t-test is a special case of linear regression.*
- If there are two groups (treated vs. control), define the dummy variable $x_i \in \{0, 1\}$. Then

$$Y_i = \beta_0 + \beta_1\, x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

- Under this setup,

$$\text{if } x_i = 0 \quad Y_i = \beta_0 + \varepsilon_i, \qquad \text{if } x_i = 1 \quad Y_i = \beta_0 + \beta_1 + \varepsilon_i.$$

- That matches the two-sample model

$$Y_i \sim \mathcal{N}(\beta_0, \sigma^2) \quad \text{or} \quad Y_i \sim \mathcal{N}(\beta_0 + \beta_1, \sigma^2),$$

  respectively. Hence, $\beta_1$ is the difference in means.
- **Testing** $H_0 : \beta_1 = 0$ in regression $\longleftrightarrow$ $H_0 : \mu_C = \mu_T$ in the two-sample $t$-test.

## 3.2 Model vs. Data (Probability vs. Statistics)

- Early in the course, we are focusing on the **model** side, i.e. specifying how $Y_i$ is *distributed* (its mean, variance, dependence on $X_i$, etc.), without yet diving into *estimators* or explicit data-based formulas.
- "Probability" is about specifying these distributions; "statistics" is about using sample data to estimate or test parameters.
- The big step forward is to see that *once you can write* $Y_i = \beta_0 + \beta_1\, x_i + \varepsilon_i$, you can carry out the usual T-test ideas (for a difference in means) by simply testing if $\beta_1 = 0$.

### 3.3 Generalizing Beyond Two Groups (Motivation for More Regression)

- The instructor emphasized that while a dummy variable captures *two* categories, we can let $x_i$ be any real value: $\{x_i \in \mathbb{R}\}$.
- Then $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ can capture many relationships: e.g. $X =$ (SAT score), $Y =$ (college GPA). Testing $\beta_1 = 0$ asks "Does GPA systematically change with SAT score?"
- This leads to the broad power of **linear regression**—it can incorporate continuous or categorical features $x_i$, or multiple features simultaneously (multiple regression).

### 3.4 Example: Automatic vs. Manual Cars (Why Regression Matters)

- Instructor gave an example of comparing fuel efficiency (miles per gallon, MPG) in cars with either *automatic* or *manual* transmissions.
- A simple two-sample $t$-test of MPG vs. transmission type might show a large, statistically significant difference. For instance, old data might show manual cars have higher MPG on average.
- However, many *other* variables affect MPG: engine size, weight, etc. Manual cars often also have smaller engines (lower weight), so "Manual vs. Automatic" alone might be conflating the effect of engine size with the transmission type.
- **Multiple regression** can adjust for these confounding variables (engine size, weight, etc.), allowing one to isolate the effect of transmission itself.

### 3.5 Key Takeaways

1. **Two-sample $t$-test $\longleftrightarrow$ simple linear regression** with a 0–1 indicator.
2. **Testing** $\beta_1 = 0$ in the regression is exactly testing if the mean responses in two groups are equal.
3. Linear regression extends further: it lets $x_i$ be continuous or includes multiple $x$'s (multiple regression) to handle confounding factors.

## 4 Conclusion & Next Steps

- We have shown **conceptually** and **algebraically** how the classical two-sample $t$-test is a particular case of the simple linear regression model.
- The "treatment" vs. "control" distinction is encoded by a dummy variable $x_i \in \{0, 1\}$. Under the null hypothesis $\beta_1 = 0$, we are asserting no difference in means.
- In practice, linear regression is far more flexible, especially as we move into *multiple regression*, enabling us to include extra predictors, account for confounding, and refine our understanding of how variables relate to each other.

*End of compiled lecture notes.*