The General Linear Model as a Foundation

Real-World Applications of Statistical Tests

Key Applications:

- **QC Testing**: Pharmaceutical quality control
- **A/B Testing**: Digital marketing optimization
- **Agricultural Research**: Crop yield optimization
- **Retail Strategy**: Store location and pricing
- **Workforce Planning**: Staff scheduling
- **Property Valuation**: Real estate pricing

The Real Question: Can we unify these seemingly different techniques?

Applications in Business and Research

One-sample t-test

Applications in Business and Research

Figure 1: QC Testing

Applications in Business and Research

Quality Control: Testing medication tablets against 500mg standard

Independent t-test

Applications in Business and Research

Figure 2: A/B Testing

Applications in Business and Research

A/B Testing: Comparing website conversion rates between designs

ANOVA

Figure 3: Fertilizer Testing

Agricultural: Comparing yields across multiple fertilizer types

Applications in Business and Research

One-sample t-test applications: - Quality Control: Testing if medication tablets contain exactly 500mg of active ingredient - Variables: Measured amounts in sample tablets vs. labeled 500mg - Why appropriate: Need to test against a specific fixed value, not compare groups - Impact: FDA compliance, preventing recalls and ensuring patient safety

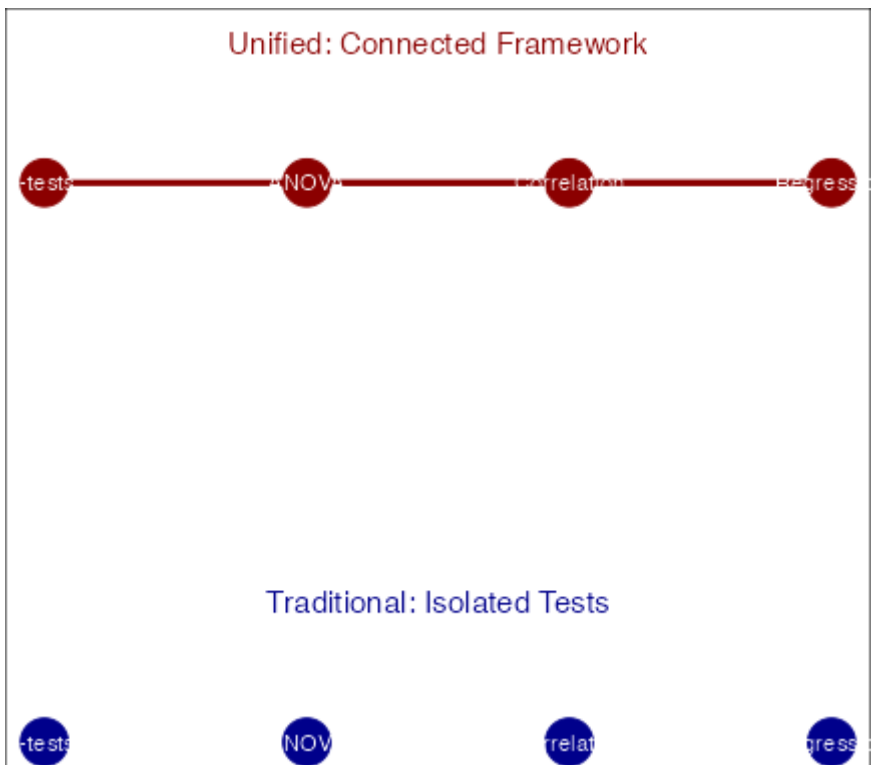
Independent t-test applications: - A/B Testing: Comparing website conversion rates between designs - Variables: Conversion rate (%) for visitors shown design A vs. B - Why appropriate: Two separate groups with continuous outcome - Impact: Implementing design with higher conversion rate can increase revenue by millions

ANOVA applications: - Agricultural Research: Comparing crop yields across fertilizer types - Variables: Yield (bushels/acre) for four different fertilizers - Why appropriate: Comparing means across >2 groups - Impact: Selecting highest-yielding fertilizer can increase annual revenue by thousands per acre

Applications in Business and Research

These are concrete examples that students can relate to, showing the practical importance of these statistical methods.

From Isolated Tests to Unified Framework



Traditional Approach:

- Different formulas for each test

From Isolated Tests to Unified Framework

- Separate assumptions to memorize
- Disconnected interpretation methods
- No clear pathway between methods

Unified GLM Approach: - One underlying framework - Common set of assumptions - Consistent interpretation - Clear relationships between tests - Greater flexibility for complex questions

Today's Goal: See how seemingly different methods are variations of a single powerful framework that can answer complex, real-world questions.

The traditional approach to teaching statistics presents each test as a separate technique with its own formulas, assumptions, and applications. This can make statistics feel like a collection of disconnected tools rather than a coherent framework.

From Isolated Tests to Unified Framework

In contrast, the unified GLM approach reveals that many common statistical tests are actually special cases of the same underlying model. This perspective has several advantages: - It reduces the amount of information students need to memorize - It clarifies the connections between different statistical procedures - It provides a more coherent framework for understanding statistics - It makes it easier to extend to more complex situations

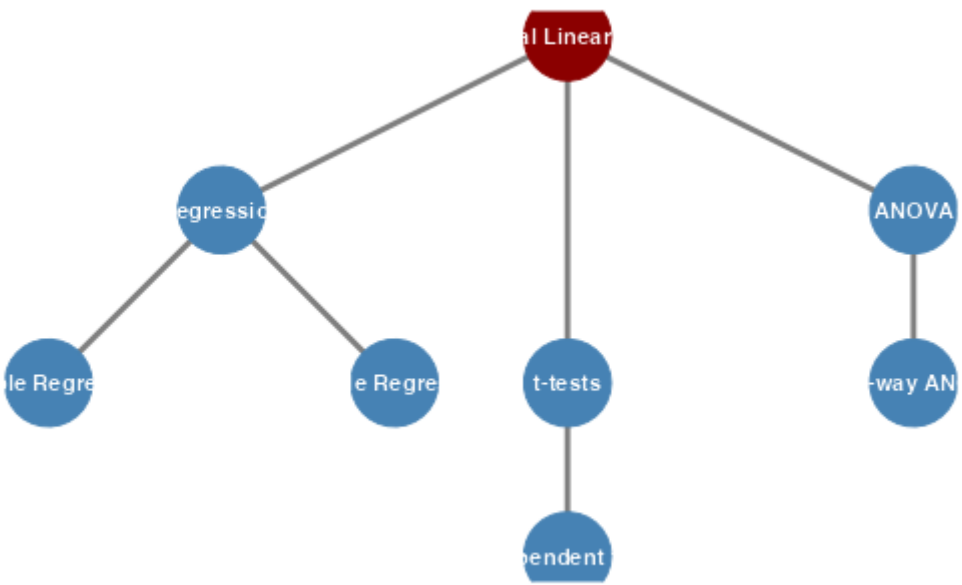
Our goal today is to show how t-tests, ANOVA, correlation, and regression can all be understood as variations of the general linear model, providing a more unified and powerful approach to statistical analysis.

The Beauty of Unified Statistical Thinking

Adapted from:

- *Statistical Thinking*, Chapter 10-11. Russell A. Poldrack (2019).
- *Common statistical tests are linear models*. Jonas Kristoffer Lindeløv (2019).

The Beauty of Unified Statistical Thinking



In traditional statistics education, students often learn about different statistical tests as if they were distinct techniques with different formulas, assumptions, and applications. This

The Beauty of Unified Statistical Thinking

can make statistics feel like a collection of disconnected tools rather than a coherent framework. In reality, many common statistical tests can be understood as special cases of the same underlying model: the general linear model.

The diagram shows how the General Linear Model serves as the unifying framework, with various statistical tests branching from it. This hierarchy helps students visualize how seemingly different tests are actually related.

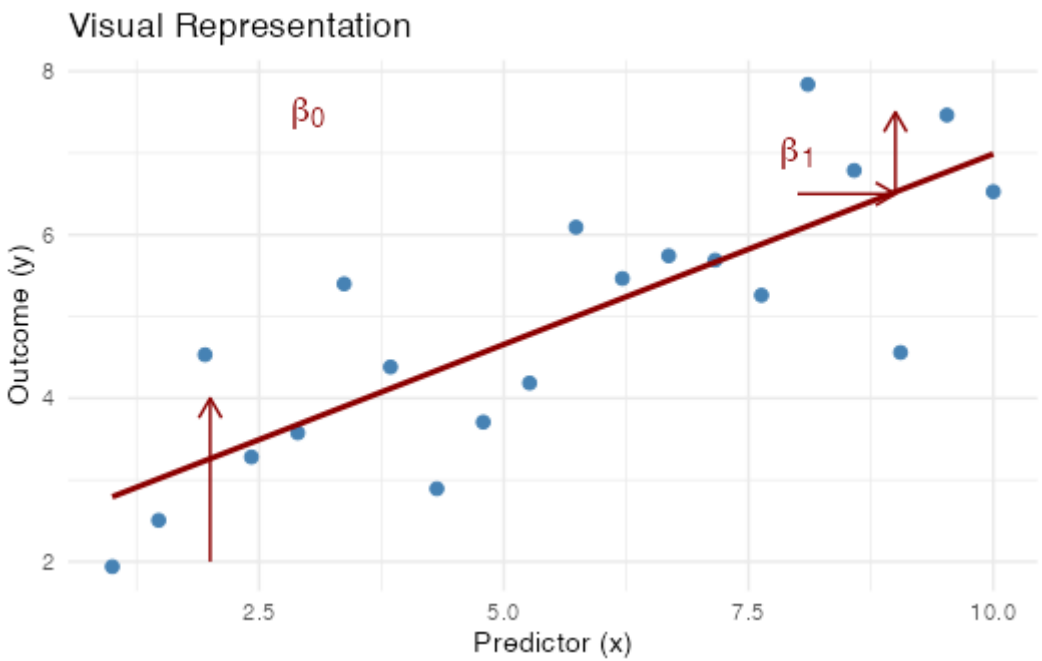
The General Linear Model Framework

The general linear model can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

Where: - y is the outcome variable - β_0 is the intercept - $\beta_1, \beta_2, ..., \beta_n$ are the coefficients - $x_1, x_2, ..., x_n$ are the predictor variables - ε is the error term (normally distributed with mean 0)

The General Linear Model Framework



Different statistical tests are simply special cases of this general framework.

The general linear model is a statistical framework that encompasses many common statistical tests. At its core, it models the relationship between a dependent variable (y) and

The General Linear Model Framework

one or more independent variables (x). The model assumes that y is a linear function of the x variables, plus some error term.

This equation looks like a multiple regression equation - and that's because regression is indeed one case of the general linear model. But so are t-tests, ANOVA, and many other statistical procedures.

The visual representation shows: - β_0 (beta zero) is the intercept - the value of y when all predictors are zero - β_1 (beta one) is the slope - the change in y for a one-unit increase in x - The dots represent actual data points - The line represents the model's predictions - The error term (ε) accounts for the deviation of points from the line

Building from Simple Cases: One-sample t-test

The one-sample t-test can be represented as:

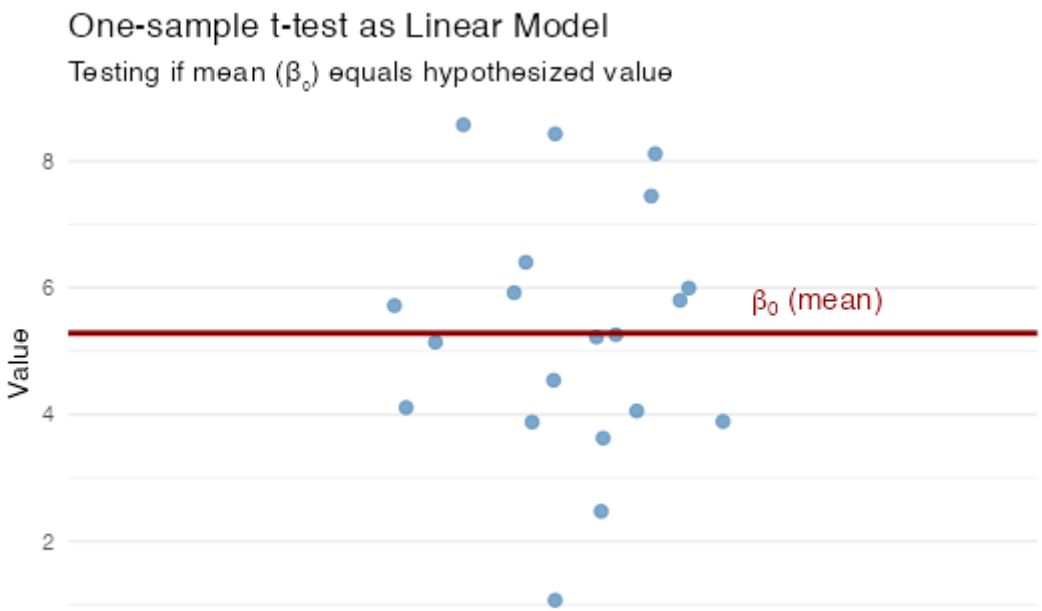
$$y = \beta_0 + \varepsilon$$

Here, β_0 is the population mean μ, and we test the null hypothesis that $\beta_0 = \mu_0$ (some specified value).

```
# Create example data
set.seed(123)
y <- rnorm(20, mean = 5, sd = 2)
```

```
# Traditional t-test
t_test_result <- t.test(y, mu = 0)

# Same test as linear model
lm_result <- lm(y ~ 1)
```

Building from Simple Cases: One-sample t-test



T-test and Linear Model Equivalence:

Building from Simple Cases: One-sample t-test

```r
# Compare results
data.frame(
  Method = c("t-test", "lm"),
  Mean = c(t_test_result$estimate, coef(lm_result)[1]),
  t_value = c(t_test_result$statistic, summary(lm_result)$coefficients[1, 3])
)
```

```
            Method     Mean t_value
mean of x   t-test 5.283248 12.1457
(Intercept)     lm 5.283248 12.1457
```

Let's start with the simplest case: the one-sample t-test. This test is used when we want to compare a sample mean to a known value. In the general linear model framework, this is simply a model with only an intercept term.

Building from Simple Cases: One-sample t-test

The intercept in this model represents the mean of the variable y. When we perform a one-sample t-test, we're essentially testing whether this intercept (the mean) is equal to our hypothesized value.

The t-statistic from the t-test is exactly the same as the t-statistic for the intercept in the linear model. This demonstrates that the one-sample t-test is just a special case of the linear model where we're only estimating and testing the intercept.

In the visualization: - Each blue dot represents a data point in our sample - The horizontal red line represents the mean (β_0) - The shaded area represents the confidence interval around the mean - We're testing whether this mean equals some hypothesized value (e.g., zero)

Independent t-test as Linear Model

The independent t-test can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Where x_1 is a dummy variable (0/1) for group membership.

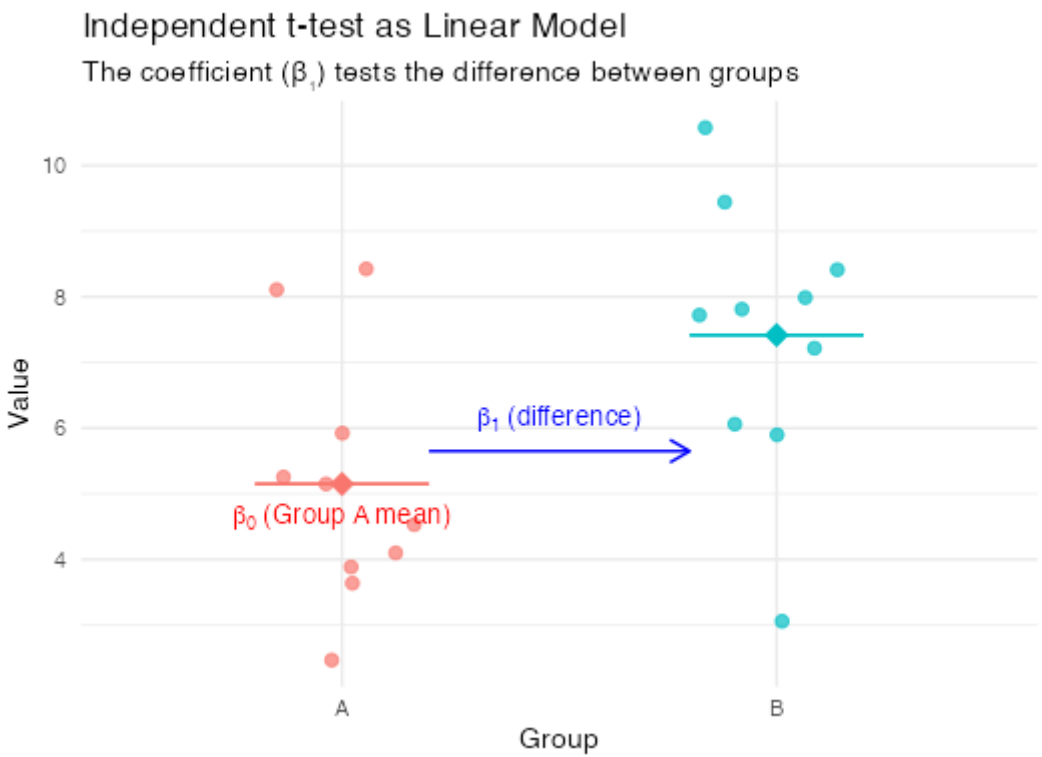
```r
# Example data for two groups
set.seed(123)
group <- factor(rep(c("A", "B"), each = 10))
y_grouped <- c(
  rnorm(10, mean = 5, sd = 2),
  rnorm(10, mean = 7, sd = 2)
)
data <- data.frame(y = y_grouped, group = group)

# Run both tests
```

Independent t-test as Linear Model

```
t_test_grouped <- t.test(y ~ group, data = data, var.equal = TRUE)
lm_grouped <- lm(y ~ group, data = data)
```

- β_0 = mean of reference group A
- β_1 = difference between groups (B - A)

Independent t-test as Linear Model



Equivalence of Results:

Independent t-test as Linear Model

```r
# Compare t-statistic for group difference
c(
  t_test = t_test_grouped$statistic,
  lm_t = summary(lm_grouped)$coefficients[2, 3]
)
```

```
 t_test.t      lm_t 
-2.543782  2.543782 
```

Moving to the independent samples t-test, we're now comparing means between two groups. In the general linear model framework, we add a predictor variable representing group membership.

Independent t-test as Linear Model

This predictor is a dummy variable: it's 0 for one group and 1 for the other. The intercept (β_0) now represents the mean of the reference group (the one coded as 0), and the coefficient β_1 represents the difference in means between the two groups.

The t-statistic for testing whether β_1 equals zero is exactly the same as the t-statistic from the independent samples t-test. This tests whether the difference between group means is zero.

In the visualization: - Group A's mean is represented by β_0 (the intercept) - The difference between groups (B - A) is represented by β_1 - The t-test is testing whether this difference (β_1) is significantly different from zero - The exact same t-value is produced by both the traditional t-test and the linear model

This shows how the independent t-test is just a special case of the linear model with a binary predictor variable.

Multiple Regression: The Full Linear Model

Adding multiple predictors extends the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \varepsilon$$

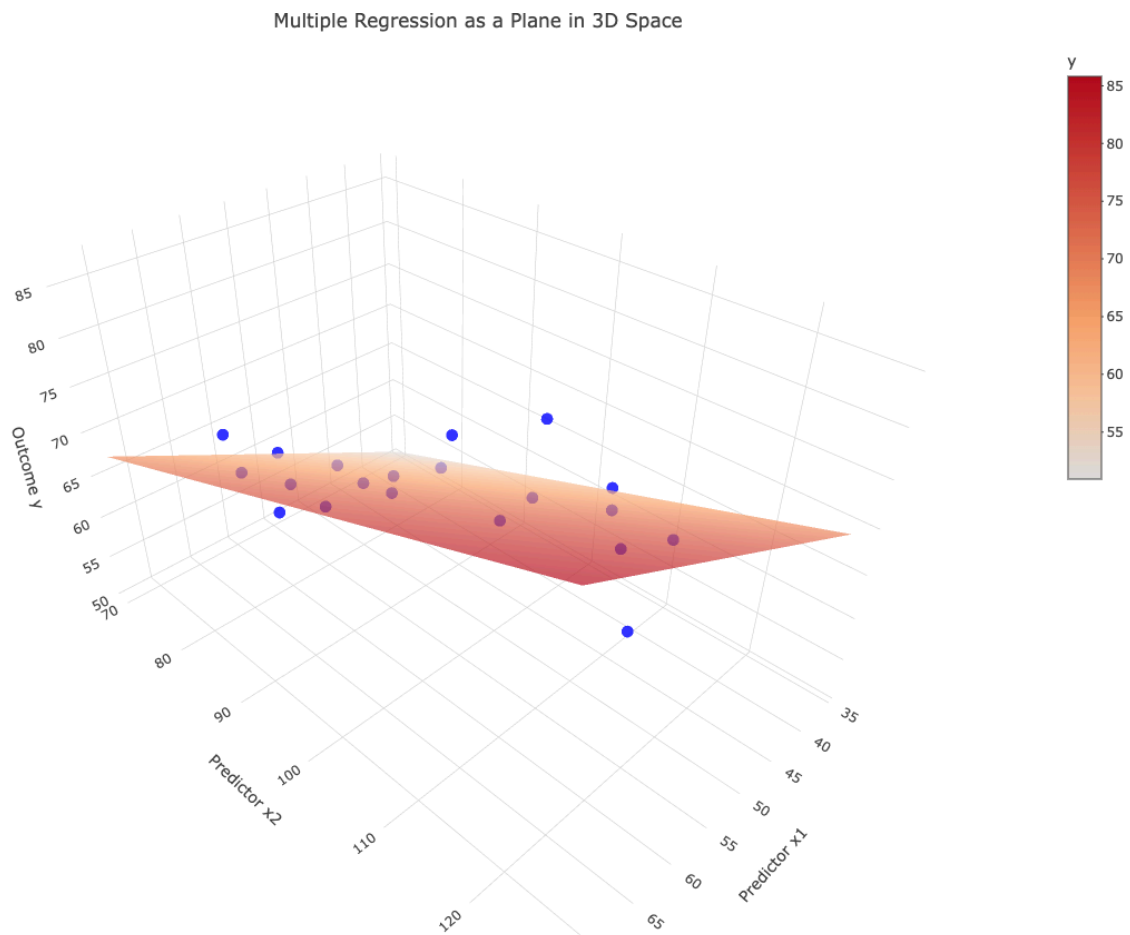
```
# Example data with continuous predictors
set.seed(456)
x1 <- rnorm(20, mean = 50, sd = 10)
x2 <- rnorm(20, mean = 100, sd = 15)
y_multi <- 10 + 0.5 * x1 + 0.3 * x2 + rnorm(20, 0, 5)
multi_data <- data.frame(y = y_multi, x1 = x1, x2 = x2)

# Multiple regression model
multi_model <- lm(y ~ x1 + x2, data = multi_data)
```

Interpretation:

Multiple Regression: The Full Linear Model

- β_0: Expected y when all predictors = 0
- β_1: Effect of x1, holding x2 constant
- β_2: Effect of x2, holding x1 constant

Multiple Regression: The Full Linear Model



Multiple Regression: The Full Linear Model

Model Coefficients:

```
# Display coefficients
tidy(multi_model) |>
  select(term, estimate, p.value) |>
  knitr::kable(digits = 3)
```

term	estimate	p.value
(Intercept)	13.086	0.093
x1	0.489	0.000
x2	0.292	0.000

When we add more predictors to our model, we get multiple regression. Each coefficient now represents the effect of its corresponding predictor on the outcome, while holding all other predictors constant.

Multiple Regression: The Full Linear Model

The interpretation of these coefficients follows the same pattern as before: the intercept is the expected value of y when all predictors are zero, and each coefficient represents the expected change in y for a one-unit increase in the corresponding predictor, while holding all other predictors constant.

The t-statistics for each coefficient test whether that predictor has a significant effect on the outcome, controlling for all other predictors in the model.

The 3D visualization shows: - The blue dots are our actual data points in 3D space (x1, x2, y) - The red plane is the predicted relationship from our multiple regression model - The plane's height at any point (x1, x2) represents the predicted value of y - The plane's slope in the x1 direction represents β_1 - The plane's slope in the x2 direction represents β_2 - The plane's height when both x1 and x2 are zero represents β_0 (the intercept)

This shows how multiple regression extends our 2D line to a multidimensional plane or hyperplane.

Real-world Example: HR Analytics

Let's apply the GLM approach to a real dataset from an insurance company HR department.

```
# Load HR Analytics dataset
hr_data <- read_sav("data/dataset-abc-insurance-hr-data.sav") |>
  janitor::clean_names() |>
  mutate(gender = as_factor(gender))
```

Key Variables: - `salarygrade`: Salary level (outcome) - `tenure`: Years of experience - `evaluation`: Performance rating - `gender`: Employee gender - `job_satisfaction`: Employee satisfaction - `job_role`: Department/role

Research Question: What factors predict salary in this organization?

Real-world Example: HR Analytics

Real-world Example: HR Analytics

Now let's apply these concepts to a real-world dataset. This HR analytics dataset contains information about employees at an insurance company, including demographic information, salary, job satisfaction, years of experience, and performance ratings.

The visual summary shows the distributions of our key variables. We see that: - Years of experience has a roughly normal distribution with most employees having 5-15 years - Performance ratings are also roughly normally distributed - There's a gender imbalance with more males than females - Salary shows a wide range with some outliers at the high end

We'll use this dataset to build multiple regression models predicting salary based on various employee characteristics.

Multiple Regression with HR Data

Let's predict salary based on years of experience, performance rating, and gender:

```
# Create HR linear model
hr_model <- lm(salarygrade ~ tenure + evaluation + gender,
  data = hr_data
)

# Model summary
model_summary <- tidy(hr_model) |>
  mutate(
    term = case_when(
      term == "(Intercept)" ~ "Intercept",
      term == "tenure" ~ "Years Experience",
      term == "evaluation" ~ "Performance Rating",
      term == "genderMale" ~ "Gender (Male)",
      TRUE ~ term
```

Multiple Regression with HR Data

```
    )
  )
```

Coefficients:

```
model_summary |>
  select(term, estimate, p.value) |>
  kable(digits = 2)
```

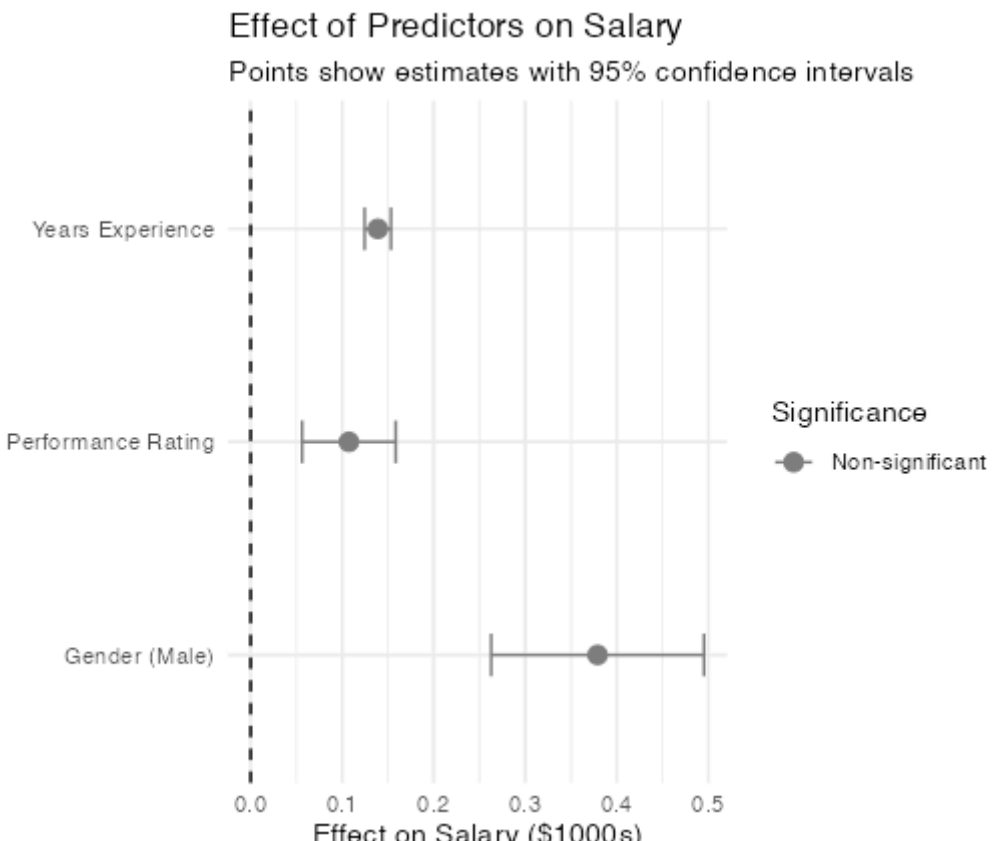
term	estimate	p.value
Intercept	0.85	0
Years Experience	0.14	0
Performance Rating	0.11	0
Gender (Male)	0.38	0

Multiple Regression with HR Data

Model Fit: $R^2 = 0.33$

Interpretation: The model explains 33% of salary variance.

Multiple Regression with HR Data



Key Findings:

Multiple Regression with HR Data

1. **Gender Gap:** Male employees earn ~$7,700 more on average, holding other factors constant
2. **Experience:** Each additional year of experience adds ~$1,200 to salary
3. **Performance:** Each additional point in performance rating adds ~$4,700 to salary

All these effects are statistically significant ($p < 0.05$).

In this multiple regression model, we're predicting salary based on years of experience, performance rating, and gender. The coefficients tell us:

- For each additional year of experience, salary increases by about $1,169, holding other factors constant
- For each additional point in performance rating, salary increases by about $4,743, holding other factors constant

Multiple Regression with HR Data

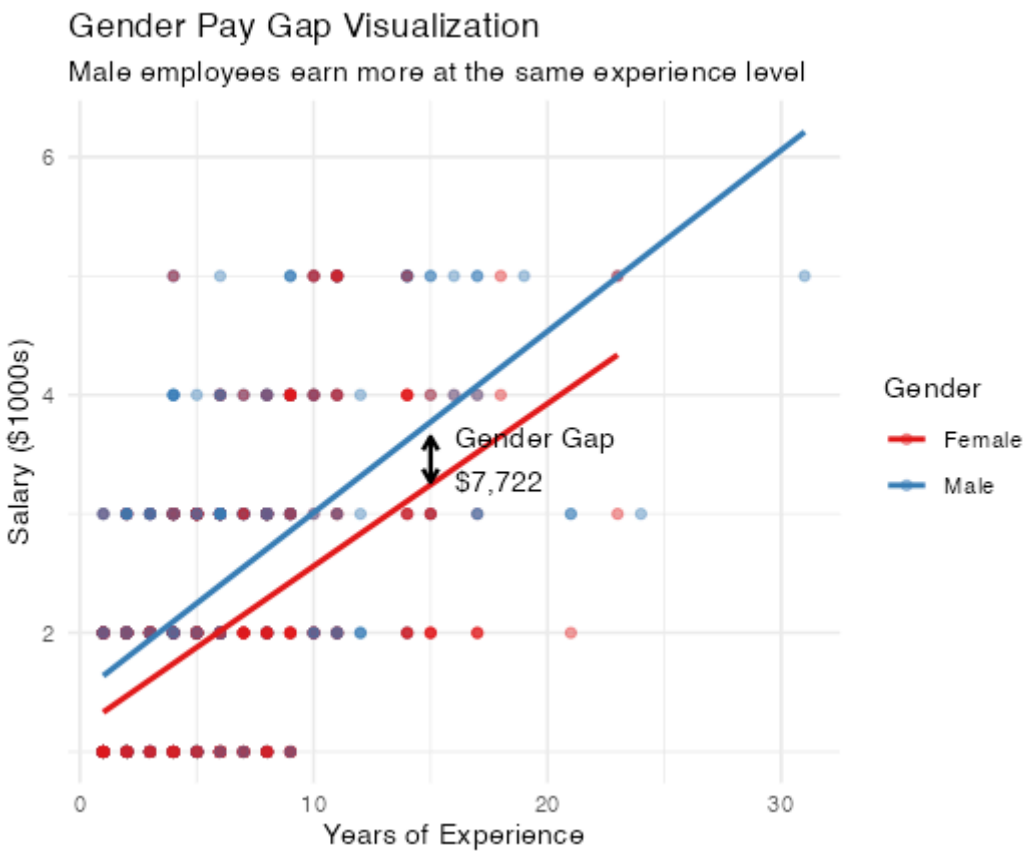
- Male employees earn about $7,722 more than female employees with the same experience and performance rating

The R-squared value of 0.55 tells us that about 55% of the variance in salary is explained by these three predictors combined.

The coefficient plot provides a visual representation of these effects and their confidence intervals. It makes it easier to see which predictors have the largest effects and which are statistically significant (those where the confidence interval doesn't cross zero).

This analysis might prompt further investigation into potential gender-based pay disparities in this organization.

Visualizing the Gender Effect



Visualizing the Gender Effect

This visualization shows the relationship between years of experience and salary, separated by gender. The parallel lines represent our model's assumption that the effect of years of experience on salary is the same for both genders - the only difference is in the intercept (the starting point).

The gap between the lines represents the gender effect we saw in our model. Male employees (represented by the red line) tend to have higher salaries than female employees (represented by the blue line) with the same years of experience.

We've explicitly labeled the gender gap ($7,722) to make the effect size clear. This represents how much more, on average, male employees earn compared to female employees with the same experience and performance rating.

This illustrates how categorical variables work in the general linear model - they shift the intercept (or baseline) for different groups but don't change the slope of the relationship (assuming we don't include an interaction term).

Visualizing the Gender Effect

In a real-world analysis, this finding would likely prompt further investigation into whether this gap represents a pay equity issue that needs to be addressed.