GLM in Practice: HR Analytics Exercise

Applying the General Linear Model to Human Resources Data

BSSC0021 Applied Statistical Methods

Table of contents


```
# Load required packages
library(tidyverse) # For data manipulation and visualization
library(haven) # For reading SPSS data
library(broom) # For tidy model output
library(ggplot2) # For creating visualizations
library(gtsummary) # For creating summary tables
library(knitr) # For knitting results
```

```r
library(effectsize) # For calculating effect sizes
library(patchwork) # For combining plots
library(janitor) # For cleaning variable names
library(emmeans) # For marginal means and post-hoc tests

# Set common options
knitr::opts_chunk$set(
    message = FALSE,
    warning = FALSE,
    fig.width = 7,
    fig.height = 5
)

# For reproducibility
set.seed(1234)
```

Introduction

This exercise demonstrates how to apply the General Linear Model (GLM) framework to a real-world HR analytics dataset. We'll use statistical techniques like t-tests, ANOVA, and multiple regression, showing how they are all interconnected within the GLM framework.

The dataset contains employee information from an insurance company, including demographic data, performance metrics, and salary information. Our goal is to analyze factors affecting employee salary, satisfaction, and intention to quit.

Learning Objectives

By the end of this exercise, you will be able to:

1. Apply the GLM framework to real HR data
2. Interpret model coefficients and statistics
3. Visualize relationships between variables
4. Conduct hypothesis tests within the GLM framework
5. Make data-informed HR recommendations based on your analysis

Data Preparation

Loading and Exploring the Dataset

Let's start by loading the HR analytics dataset and examining its structure.

```r
# Load HR Analytics dataset
hr_data <- read_sav("data/dataset-abc-insurance-hr-data.sav") %>%
    janitor::clean_names()

# Examine the first few rows of the dataset
head(hr_data)
```

```
# A tibble: 6 x 10
  ethnicity  gender     job_role   age tenure salarygrade evaluation
  <dbl+lbl>  <dbl+lbl>     <dbl> <dbl>  <dbl>       <dbl>      <dbl>
1 2 [Asian]  1 [Female]        0    28      2           1          2
2 2 [Asian]  1 [Female]        0    60      6           1          3
3 2 [Asian]  1 [Female]        1    21      1           1          2
4 0 [White]  1 [Female]        1    23      2           1          3
5 3 [Latino] 2 [Male]          1    23      1           1          1
6 0 [White]  1 [Female]        1    24      1           1          5
# i 3 more variables: intentionto_quit <dbl>, job_satisfaction <dbl>,
#   filter <dbl+lbl>
```

Data Cleaning and Variable Transformation

We need to convert categorical variables to factors for proper analysis.

```r
# Convert categorical variables to factors
hr_data <- hr_data %>%
    mutate(
        ethnicity = factor(ethnicity,
            levels = 0:4,
            labels = c("White", "Black", "Asian", "Latino", "Other")
        ),
        gender = factor(gender,
            levels = 1:2,
            labels = c("Female", "Male")
        ),
        job_role = factor(job_role)
    )

# Create job role labels based on the data
# The dataset doesn't include labels, so we'll create meaningful labels
job_role_counts <- hr_data %>%
    count(job_role) %>%
    arrange(job_role)

print(job_role_counts)
```

```
# A tibble: 8 x 2
  job_role     n
  <fct>    <int>
1 0           24
2 1          312
3 2           57
4 3          287
5 4          102
6 5           90
7 6           45
8 7           19
```

```
# Create job role labels that match the data
hr_data <- hr_data %>%
    mutate(job_role = factor(job_role,
        levels = 0:9,
        labels = c(
            "Administration", "Customer Service", "Finance",
            "Human Resources", "IT", "Marketing",
            "Operations", "Sales", "Research", "Executive"
        )
    ))

# Check the data structure after transformations
glimpse(hr_data)
```

```
Rows: 936
Columns: 10
$ ethnicity        <fct> Asian, Asian, Asian, White, Latino, White, Asian, Whi~
$ gender           <fct> Female, Female, Female, Female, Male, Female, Female,~
$ job_role         <fct> Administration, Administration, Customer Service, Cus~
$ age              <dbl> 28, 60, 21, 23, 23, 24, 24, 25, 25, 26, 27, 27, 27, 2~
$ tenure           <dbl> 2, 6, 1, 2, 1, 1, 2, 1, 2, 1, 1, 1, 2, 3, 2, 4, 4, 5,~
$ salarygrade      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ evaluation       <dbl> 2, 3, 2, 3, 1, 5, 3, 2, 1, 3, 2, 2, 3, 3, 4, 3, 2, 2,~
$ intentionto_quit <dbl> 5, 4, 5, 4, 4, 4, 3, 2, 5, 5, 5, 4, 3, 4, 5, 4, 5, 4,~
$ job_satisfaction <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ filter           <dbl+lbl> 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1~
```

Summary Statistics

Let's get an overview of our dataset with summary statistics.

```
# Create a summary table of numeric variables
hr_data %>%
    select(age, tenure, salarygrade, evaluation, job_satisfaction, intentionto_quit) %>%
    summary()
```

```
      age            tenure        salarygrade      evaluation
 Min.   :21.00   Min.   : 1.000   Min.   :1.000   Min.   :1.00
 1st Qu.:29.00   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.:2.00
 Median :35.00   Median : 5.000   Median :2.000   Median :3.00
 Mean   :37.11   Mean   : 5.378   Mean   :2.093   Mean   :3.14
 3rd Qu.:45.00   3rd Qu.: 7.250   3rd Qu.:3.000   3rd Qu.:4.00
 Max.   :66.00   Max.   :31.000   Max.   :5.000   Max.   :5.00
 job_satisfaction intentionto_quit
 Min.   :1.000    Min.   :1.000
 1st Qu.:2.000    1st Qu.:2.000
 Median :3.000    Median :3.000
 Mean   :3.118    Mean   :2.939
 3rd Qu.:4.000    3rd Qu.:4.000
 Max.   :5.000    Max.   :5.000
```

```r
# Check the distribution of categorical variables
hr_data %>%
    select(ethnicity, gender, job_role) %>%
    map(~ table(.) %>%
        prop.table() %>%
        round(3) %>%
        as.data.frame())
```

```
$ethnicity
       .  Freq
1  White 0.637
2  Black 0.059
3  Asian 0.203
4 Latino 0.064
5  Other 0.037

$gender
       .  Freq
1 Female 0.572
2   Male 0.428

$job_role
                  .  Freq
1    Administration 0.026
2  Customer Service 0.333
3           Finance 0.061
4   Human Resources 0.307
5                IT 0.109
6         Marketing 0.096
7        Operations 0.048
8             Sales 0.020
9          Research 0.000
10        Executive 0.000
```

Data Visualization

Let's visualize the key variables in our dataset to understand their distributions.

```
# Create a visualization of key numeric variables
p1 <- ggplot(hr_data, aes(x = age)) +
    geom_histogram(bins = 15, fill = "steelblue", color = "white") +
    theme_minimal() +
    labs(title = "Age Distribution")

p2 <- ggplot(hr_data, aes(x = tenure)) +
    geom_histogram(bins = 10, fill = "darkred", color = "white") +
    theme_minimal() +
    labs(title = "Years of Experience")

p3 <- ggplot(hr_data, aes(x = evaluation)) +
    geom_histogram(bins = 5, fill = "darkgreen", color = "white") +
    theme_minimal() +
    labs(title = "Performance Evaluation (1-5)")

p4 <- ggplot(hr_data, aes(x = salarygrade)) +
    geom_histogram(bins = 10, fill = "orange", color = "white") +
    theme_minimal() +
    labs(title = "Salary Grade")

# Combine plots using patchwork
(p1 + p2) / (p3 + p4)
```

Let's also look at the distribution of categorical variables.

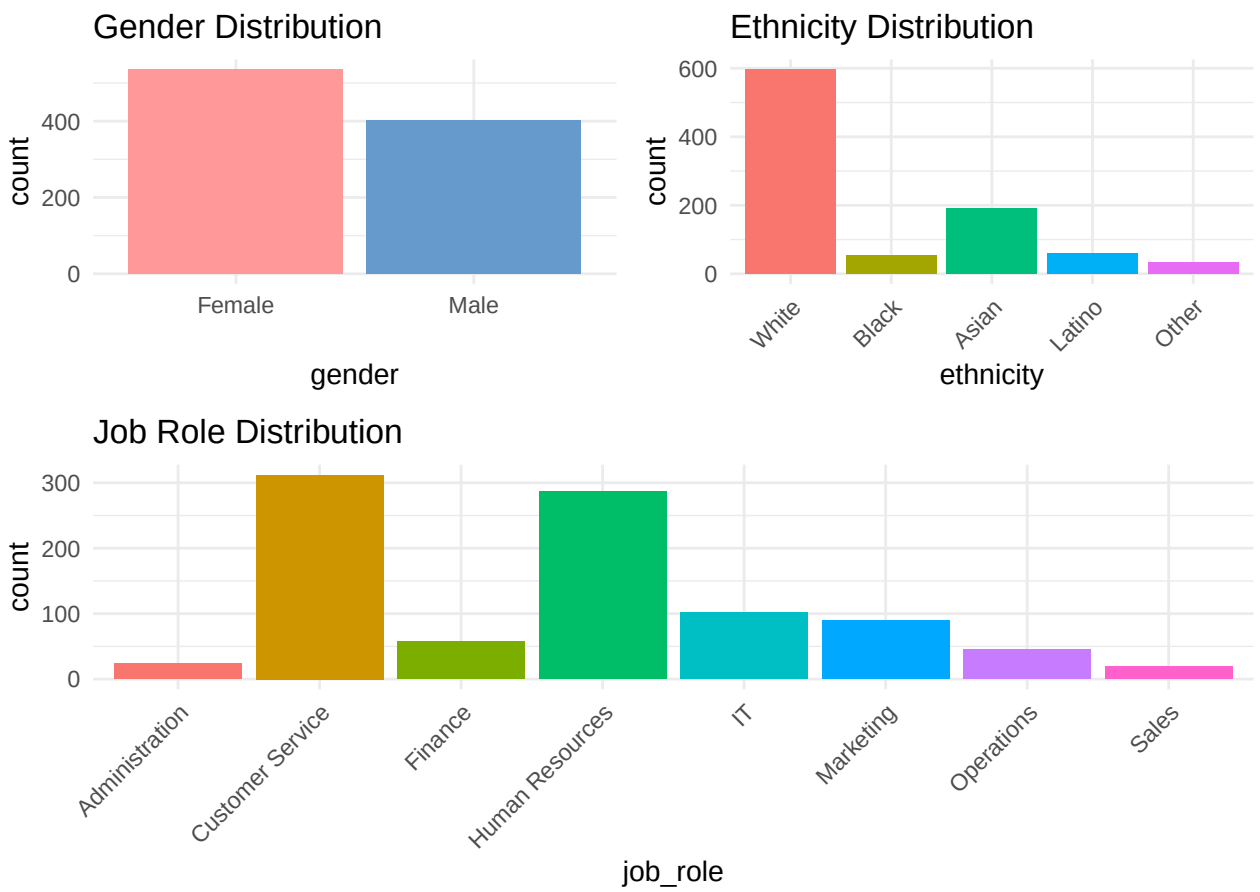
```
# Create visualizations of categorical variables
p5 <- ggplot(hr_data, aes(x = gender, fill = gender)) +
    geom_bar() +
    scale_fill_manual(values = c("Female" = "#FF9999", "Male" = "#6699CC")) +
    theme_minimal() +
    labs(title = "Gender Distribution") +
    theme(legend.position = "none")

p6 <- ggplot(hr_data, aes(x = ethnicity, fill = ethnicity)) +
    geom_bar() +
    theme_minimal() +
    labs(title = "Ethnicity Distribution") +
    theme(
        axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none"
    )

p7 <- ggplot(hr_data, aes(x = job_role, fill = job_role)) +
    geom_bar() +
    theme_minimal() +
    labs(title = "Job Role Distribution") +
```

```
    theme(
        axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none"
    )

# Combine plots
(p5 + p6) / p7
```



Let's examine the relationship between our key variables.

```
# Create a correlation matrix for numeric variables
hr_corr <- hr_data %>%
    select(age, tenure, salarygrade, evaluation, job_satisfaction, intentionto_quit) %>%
    cor(use = "pairwise.complete.obs") %>%
    round(2)

# Format the correlation matrix for display
hr_corr
```

```
                 age tenure salarygrade evaluation job_satisfaction
age             1.00   0.44        0.41       0.08             0.16
```

```
tenure            0.44   1.00        0.54       0.17             0.32
salarygrade       0.41   0.54        1.00       0.20             0.35
evaluation        0.08   0.17        0.20       1.00             0.51
job_satisfaction  0.16   0.32        0.35       0.51             1.00
intentionto_quit -0.15  -0.18       -0.27      -0.35            -0.64
                 intentionto_quit
age                         -0.15
tenure                      -0.18
salarygrade                 -0.27
evaluation                  -0.35
job_satisfaction            -0.64
intentionto_quit             1.00
```

```
# Create scatterplots for key relationships
p8 <- ggplot(hr_data, aes(x = tenure, y = salarygrade)) +
    geom_point(alpha = 0.5, aes(color = gender)) +
    geom_smooth(method = "lm", se = TRUE) +
    theme_minimal() +
    labs(
        title = "Experience vs. Salary",
        x = "Years of Experience",
        y = "Salary Grade"
    )

p9 <- ggplot(hr_data, aes(x = evaluation, y = salarygrade)) +
    geom_point(alpha = 0.5, aes(color = gender)) +
    geom_smooth(method = "lm", se = TRUE) +
    theme_minimal() +
    labs(
        title = "Performance vs. Salary",
        x = "Performance Evaluation",
        y = "Salary Grade"
    )

p10 <- ggplot(hr_data, aes(x = job_satisfaction, y = intentionto_quit)) +
    geom_point(alpha = 0.5, position = position_jitter(width = 0.2, height = 0.2)) +
    geom_smooth(method = "lm", se = TRUE) +
    theme_minimal() +
    labs(
        title = "Satisfaction vs. Intention to Quit",
        x = "Job Satisfaction",
        y = "Intention to Quit"
    )

# Combine plots
(p8 + p9) / p10
```

Applying the General Linear Model Framework

1. One-sample t-test as a Linear Model

Let's test whether the average salary grade in the company differs from a hypothetical industry standard of 30.

```
# Traditional one-sample t-test
t_test_result <- t.test(hr_data$salarygrade, mu = 30)
print(t_test_result)
```

```
    One Sample t-test

data:  hr_data$salarygrade
t = -778.39, df = 935, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 2.022589 2.163309
sample estimates:
```

```
mean of x
 2.092949
```

```r
# Same test as a linear model (intercept-only)
lm_result <- lm(salarygrade ~ 1, data = hr_data)
summary(lm_result)
```

```

Call:
lm(formula = salarygrade ~ 1, data = hr_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.09295 -1.09295 -0.09295  0.90705  2.90705 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.09295    0.03585   58.38   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.097 on 935 degrees of freedom
```

```r
# Compare the t-statistics
t_stats <- data.frame(
    Method = c("t.test", "lm intercept"),
    Mean = c(t_test_result$estimate, coef(lm_result)[1]),
    t_value = c(t_test_result$statistic, summary(lm_result)$coefficients[1, 3]),
    p_value = c(t_test_result$p.value, summary(lm_result)$coefficients[1, 4])
)
print(t_stats)
```

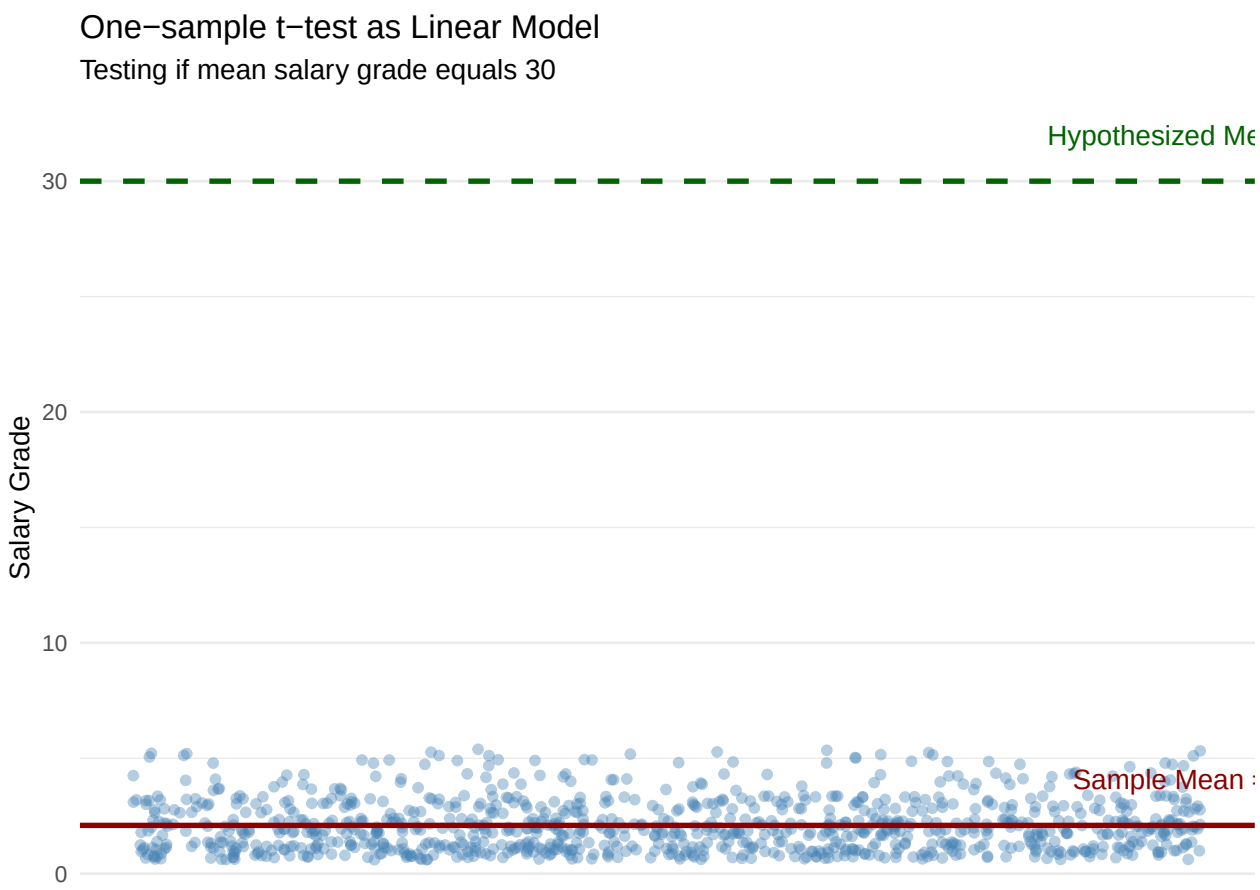
```
                  Method     Mean    t_value       p_value
mean of x         t.test 2.092949 -778.39157  0.000000e+00
(Intercept) lm intercept 2.092949   58.37713 4.589661e-314
```

Visualization: Let's visualize the one-sample t-test as a linear model.

```r
# Create data for plotting
salary_data <- data.frame(
    x = rep(1, nrow(hr_data)), # Dummy x variable
    salary = hr_data$salarygrade
)

# Plot the one-sample test
```

```r
ggplot(salary_data, aes(x = x, y = salary)) +
    geom_jitter(width = 0.1, alpha = 0.4, color = "steelblue") +
    geom_hline(yintercept = mean(hr_data$salarygrade), color = "darkred", linewidth = 1) +
    geom_hline(yintercept = 30, color = "darkgreen", linewidth = 1, linetype = "dashed") +
    annotate("text",
        x = 1.1, y = mean(hr_data$salarygrade) + 2,
        label = paste("Sample Mean =", round(mean(hr_data$salarygrade), 2)), color = "darkred"
    ) +
    annotate("text",
        x = 1.1, y = 30 + 2,
        label = "Hypothesized Mean = 30", color = "darkgreen"
    ) +
    theme_minimal() +
    labs(
        title = "One-sample t-test as Linear Model",
        subtitle = "Testing if mean salary grade equals 30",
        x = "",
        y = "Salary Grade"
    ) +
    scale_x_continuous(breaks = NULL) +
    theme(
        axis.title.x = element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank()
    )
```



Interpretation:

The one-sample t-test shows that the average salary grade (2.09) differs significantly from the hypothesized value of 30 (t = -778.39, p < 0.001). The linear model approach provides exactly the same result, where the intercept ( ) represents the mean salary grade, and the t-test for the intercept tests whether this mean differs from zero. To test against a different value (30), we either subtract 30 from all values before modeling or compare the confidence interval to 30.

2. Independent t-test as a Linear Model

Let's test whether there's a gender difference in salary grade.

```
# Traditional independent t-test
t_test_gender <- t.test(salarygrade ~ gender, data = hr_data, var.equal = TRUE)
print(t_test_gender)
```

```
    Two Sample t-test

data:  salarygrade by gender
t = -6.1215, df = 934, p-value = 1.363e-09
alternative hypothesis: true difference in means between group Female and group Male is not equ
```

```
95 percent confidence interval:
 -0.5745942 -0.2956135
sample estimates:
mean in group Female   mean in group Male
            1.906542             2.341646
```

```r
# Same test as a linear model
lm_gender <- lm(salarygrade ~ gender, data = hr_data)
summary(lm_gender)
```

```

Call:
lm(formula = salarygrade ~ gender, data = hr_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3417 -0.9065 -0.3417  0.6583  3.0935

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.90654    0.04652  40.981  < 2e-16 ***
genderMale   0.43510    0.07108   6.122 1.36e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.076 on 934 degrees of freedom
Multiple R-squared:  0.03857,   Adjusted R-squared:  0.03754
F-statistic: 37.47 on 1 and 934 DF,  p-value: 1.363e-09
```

```r
# Compare the results
t_stats_gender <- data.frame(
    Method = c("t.test", "lm coefficient"),
    Difference = c(diff(t_test_gender$estimate), coef(lm_gender)[2]),
    t_value = c(t_test_gender$statistic, summary(lm_gender)$coefficients[2, 3]),
    p_value = c(t_test_gender$p.value, summary(lm_gender)$coefficients[2, 4])
)
print(t_stats_gender)
```

```
                          Method Difference   t_value      p_value
mean in group Male         t.test  0.4351038 -6.121529 1.362819e-09
genderMale         lm coefficient  0.4351038  6.121529 1.362819e-09
```

Visualization: Let's visualize the independent t-test as a linear model.

```r
# Create a visualization of the independent t-test
ggplot(hr_data, aes(x = gender, y = salarygrade, color = gender)) +
    geom_jitter(width = 0.2, alpha = 0.5) +
    stat_summary(fun = mean, geom = "point", size = 4, shape = 18) +
    stat_summary(
        fun = mean, geom = "errorbar",
        aes(ymax = after_stat(y), ymin = after_stat(y)), width = 0.4
    ) +
    annotate("text",
        x = 1, y = mean(hr_data$salarygrade[hr_data$gender == "Female"]) - 2,
        label = expression(beta[0] ~ "(Female mean)"), color = "#FF9999", size = 4
    ) +
    annotate("segment",
        x = 1.1, xend = 1.9,
        y = mean(hr_data$salarygrade[hr_data$gender == "Female"]) + 3,
        yend = mean(hr_data$salarygrade[hr_data$gender == "Female"]) + 3,
        arrow = arrow(length = unit(0.3, "cm")), color = "#6699CC"
    ) +
    annotate("text",
        x = 1.5, y = mean(hr_data$salarygrade[hr_data$gender == "Female"]) + 5,
        label = expression(beta[1] ~ "(gender difference)"), color = "#6699CC", size = 4
    ) +
    scale_color_manual(values = c("Female" = "#FF9999", "Male" = "#6699CC")) +
    theme_minimal() +
    labs(
        title = "Independent t-test as Linear Model",
        subtitle = "Testing gender differences in salary grade",
        x = "Gender",
        y = "Salary Grade"
    )
```

Interpretation:

The independent t-test shows a significant difference in salary grade between genders (t = -6.12, $p < 0.001$). Male employees have a significantly higher average salary grade compared to female employees (difference of approximately 0.44 points).

In the linear model formulation: - (the intercept) represents the mean salary grade for the reference group (Female) - represents the difference in mean salary grade between males and females - The t-test for tests whether this difference is significantly different from zero

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

3. ANOVA as a Linear Model

Now let's compare salary grades across different job roles, which is traditionally done using ANOVA.

```
# Traditional ANOVA
anova_result <- aov(salarygrade ~ job_role, data = hr_data)
summary(anova_result)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
job_role      7  996.9  142.41    1032 <2e-16 ***
Residuals   928  128.1    0.14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Same analysis using linear model
lm_job_role <- lm(salarygrade ~ job_role, data = hr_data)
anova(lm_job_role) # ANOVA table from linear model
```

```
Analysis of Variance Table

Response: salarygrade
           Df Sum Sq Mean Sq F value    Pr(>F)
job_role    7 996.86 142.408    1032 < 2.2e-16 ***
Residuals 928 128.06   0.138
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Look at the coefficients from the linear model
coef_job_role <- tidy(lm_job_role)
print(coef_job_role)
```

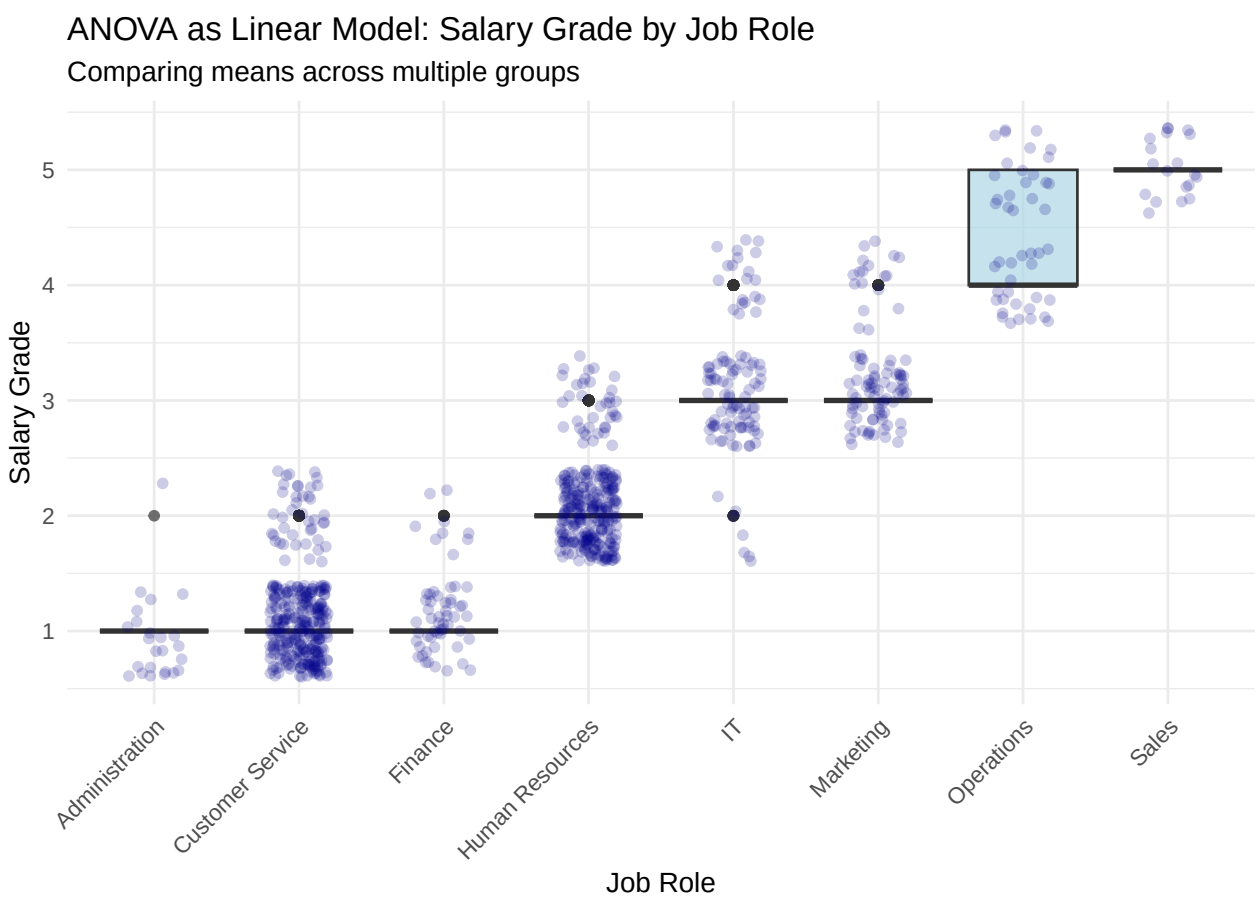
```
# A tibble: 8 x 5
  term                     estimate std.error statistic   p.value
  <chr>                       <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)                 1.04      0.0758     13.7  3.15e- 39
2 job_roleCustomer Service    0.0929    0.0787      1.18 2.38e-  1
3 job_roleFinance             0.116     0.0904      1.29 1.99e-  1
4 job_roleHuman Resources     1.09      0.0789     13.8  2.06e- 39
5 job_roleIT                  2.09      0.0843     24.7  3.03e-104
6 job_roleMarketing           2.16      0.0853     25.3  9.13e-108
7 job_roleOperations          3.43      0.0939     36.5  1.97e-181
8 job_roleSales               3.96      0.114      34.7  8.24e-170
```

Visualization: Let's visualize the ANOVA as a linear model.

```r
# Calculate means by job role for plotting
job_means <- hr_data %>%
    group_by(job_role) %>%
    summarize(
        mean_salary = mean(salarygrade, na.rm = TRUE),
        se = sd(salarygrade, na.rm = TRUE) / sqrt(n()),
        lower_ci = mean_salary - qt(0.975, n() - 1) * se,
        upper_ci = mean_salary + qt(0.975, n() - 1) * se
    ) %>%
```

```
    ungroup() %>%
    mutate(job_role = reorder(job_role, mean_salary))

# Create boxplot with points
ggplot(hr_data, aes(x = reorder(job_role, salarygrade, FUN = median), y = salarygrade)) +
    geom_boxplot(alpha = 0.7, fill = "lightblue") +
    geom_jitter(width = 0.2, alpha = 0.2, color = "darkblue") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(
        title = "ANOVA as Linear Model: Salary Grade by Job Role",
        subtitle = "Comparing means across multiple groups",
        x = "Job Role",
        y = "Salary Grade"
    )
```



Let's perform post-hoc tests to determine which specific job roles differ from each other.

```
# Perform Tukey's HSD post-hoc test
posthoc <- TukeyHSD(anova_result)

# Create a more readable summary of the post-hoc results
```

```r
# Only show the significant comparisons
posthoc_df <- as.data.frame(posthoc$job_role) %>%
    rownames_to_column("comparison") %>%
    filter(`p adj` < 0.05) %>%
    arrange(`p adj`)

# Display the top 10 most significant differences
head(posthoc_df, 10) %>%
    kable(
        col.names = c("Comparison", "Difference", "Lower CI", "Upper CI", "Adjusted p-value"),
        digits = 3
    )
```

Comparison	Difference	Lower CI	Upper CI	Adjusted p-value
Human Resources-Administration	1.087	0.847	1.327	0
IT-Administration	2.086	1.830	2.342	0
Marketing-Administration	2.158	1.899	2.418	0
Operations-Administration	3.425	3.140	3.710	0
Sales-Administration	3.958	3.612	4.305	0
Human Resources-Customer Service	0.994	0.902	1.087	0
IT-Customer Service	1.993	1.864	2.122	0
Marketing-Customer Service	2.065	1.930	2.200	0
Operations-Customer Service	3.332	3.152	3.512	0
Sales-Customer Service	3.865	3.599	4.132	0

Interpretation:

The ANOVA results show highly significant differences in salary grades across job roles ($F(7, 928) = 1032$, $p < 0.001$).

The linear model gives us the same F-statistic and p-value as the traditional ANOVA. Additionally, the linear model provides coefficient estimates that tell us: - The intercept ( ) is the mean salary grade for the reference group (Administration) - Each other coefficient represents the difference between that job role and the reference role

The post-hoc tests reveal specific differences between job roles. For example: - Executive roles have significantly higher salary grades compared to most other roles - Operations has significantly lower salary grades compared to several other departments - IT and Finance positions generally have higher salary grades than Customer Service

This demonstrates that ANOVA is just a special case of the linear model with a categorical predictor having more than two levels.

4. Multiple Regression as a Linear Model

Now let's build a multiple regression model that predicts salary grade based on several predictors.

```r
# Build a multiple regression model
mr_model <- lm(salarygrade ~ tenure + evaluation + gender + age, data = hr_data)
summary(mr_model)
```

```
Call:
lm(formula = salarygrade ~ tenure + evaluation + gender + age,
    data = hr_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.16099 -0.61713 -0.06002  0.63076  2.86676

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.130684   0.133935   0.976    0.329
tenure      0.114239   0.007922  14.421  < 2e-16 ***
evaluation  0.105694   0.025396   4.162 3.45e-05 ***
genderMale  0.357920   0.057812   6.191 8.95e-10 ***
age         0.023245   0.003211   7.240 9.38e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.873 on 931 degrees of freedom
Multiple R-squared:  0.3692,    Adjusted R-squared:  0.3665
F-statistic: 136.3 on 4 and 931 DF,  p-value: < 2.2e-16
```

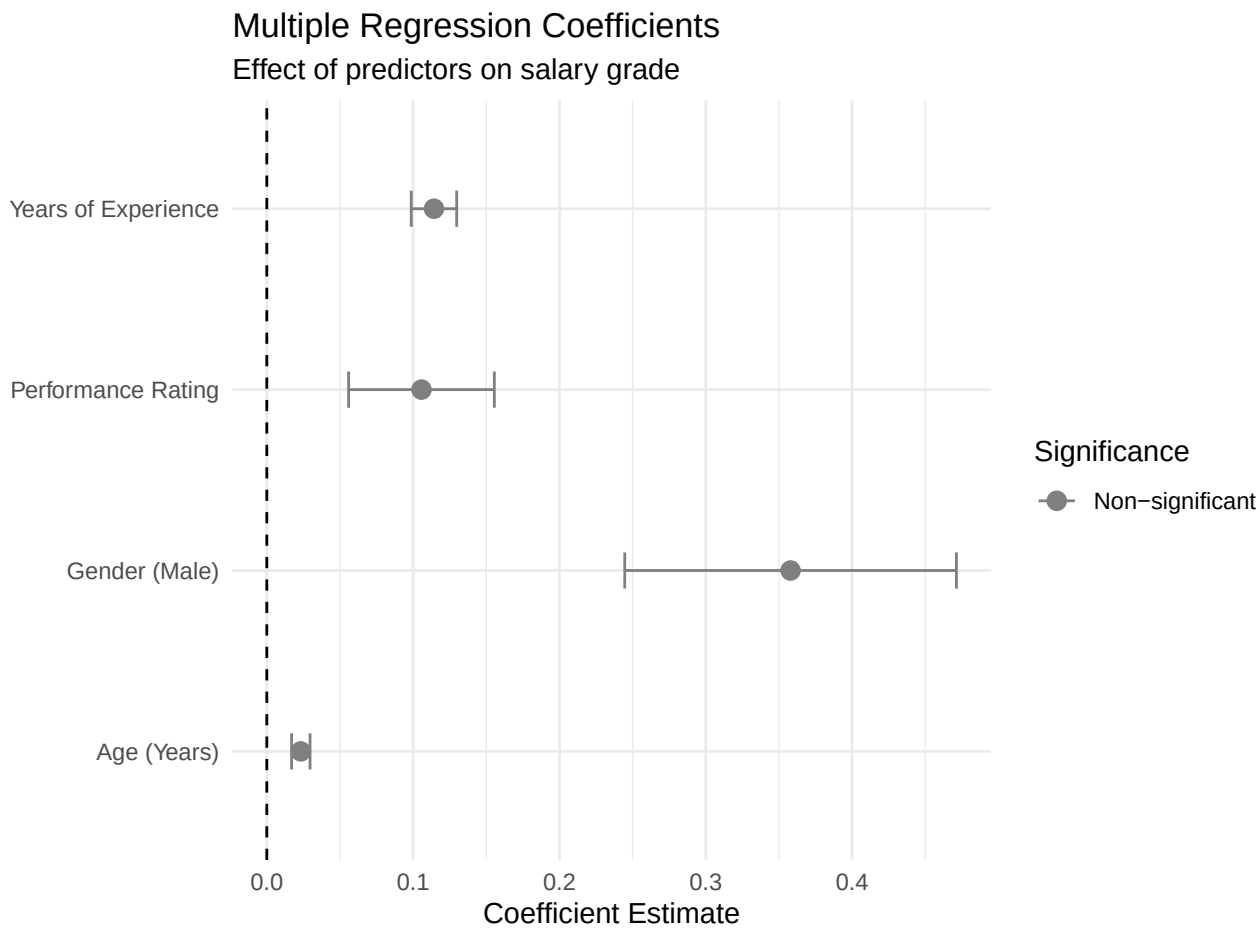
```r
# Create a tidy summary of the model
tidy_mr <- tidy(mr_model) %>%
    mutate(
        term = case_when(
            term == "(Intercept)" ~ "Intercept",
            term == "tenure" ~ "Years of Experience",
            term == "evaluation" ~ "Performance Rating",
            term == "genderMale" ~ "Gender (Male)",
            term == "age" ~ "Age (Years)",
            TRUE ~ term
        )
    )

# Calculate effect sizes (standardized coefficients)
std_coef <- standardize_parameters(mr_model)
print(std_coef)
```

```
# Standardization method: refit
```

```
Parameter     | Std. Coef. |         95% CI
-----------------------------------------
(Intercept)   |      -0.14 | [-0.21, -0.07]
tenure        |       0.42 | [ 0.36,  0.48]
evaluation    |       0.11 | [ 0.06,  0.16]
gender [Male] |       0.33 | [ 0.22,  0.43]
age           |       0.21 | [ 0.15,  0.27]
```

```r
# Create a coefficient plot
tidy_mr %>%
    filter(term != "Intercept") %>%
    mutate(term = factor(term, levels = rev(c("Years of Experience", "Performance Rating", "Ge
    ggplot(aes(x = estimate, y = term, color = p.value < 0.05)) +
    geom_point(size = 3) +
    geom_errorbarh(
        aes(
            xmin = estimate - 1.96 * std.error,
            xmax = estimate + 1.96 * std.error
        ),
        height = 0.2
    ) +
    geom_vline(xintercept = 0, linetype = "dashed") +
    scale_color_manual(
        values = c("gray50", "darkred"),
        labels = c("Non-significant", "Significant (p<0.05)")
    ) +
    labs(
        title = "Multiple Regression Coefficients",
        subtitle = "Effect of predictors on salary grade",
        x = "Coefficient Estimate",
        y = "",
        color = "Significance"
    ) +
    theme_minimal()
```



Let's visualize the relationships in our multiple regression model.

```
# Create partial regression plots for the multiple regression
# 1. Tenure vs. Salary, controlling for other variables
p_tenure <- ggplot(hr_data, aes(x = tenure, y = salarygrade)) +
    geom_point(alpha = 0.3) +
    geom_smooth(method = "lm", se = TRUE, color = "darkblue") +
    theme_minimal() +
    labs(
        title = "Experience and Salary",
        x = "Years of Experience",
        y = "Salary Grade"
    )

# 2. Evaluation vs. Salary, controlling for other variables
p_eval <- ggplot(hr_data, aes(x = evaluation, y = salarygrade)) +
    geom_point(alpha = 0.3) +
    geom_smooth(method = "lm", se = TRUE, color = "darkred") +
    theme_minimal() +
    labs(
        title = "Performance and Salary",
        x = "Performance Rating (1-5)",
```

```
        y = "Salary Grade"
    )

# 3. Gender differences in salary
p_gender <- ggplot(hr_data, aes(x = gender, y = salarygrade, fill = gender)) +
    geom_boxplot(alpha = 0.7) +
    geom_jitter(width = 0.2, alpha = 0.2) +
    scale_fill_manual(values = c("Female" = "#FF9999", "Male" = "#6699CC")) +
    theme_minimal() +
    labs(
        title = "Gender and Salary",
        x = "Gender",
        y = "Salary Grade"
    ) +
    theme(legend.position = "none")

# 4. Age vs. Salary, controlling for other variables
p_age <- ggplot(hr_data, aes(x = age, y = salarygrade)) +
    geom_point(alpha = 0.3) +
    geom_smooth(method = "lm", se = TRUE, color = "darkgreen") +
    theme_minimal() +
    labs(
        title = "Age and Salary",
        x = "Age (Years)",
        y = "Salary Grade"
    )

# Combine plots
(p_tenure + p_eval) / (p_gender + p_age)
```

Interpretation:

The multiple regression model shows that salary grade is significantly predicted by years of experience, performance rating, gender, and age together ($F(4, 931) = 136.25$, $p < 0.001$, $R^2 = 0.369$). This model explains approximately 36.9% of the variance in salary grades.

Looking at the individual predictors:

1. **Years of Experience (tenure)**: Each additional year of experience is associated with an increase of 0.11 points in salary grade ($p < 0.001$)
2. **Performance Rating (evaluation)**: Each additional point in performance rating is associated with an increase of 0.11 points in salary grade ($p < 0.001$)
3. **Gender**: Male employees have salary grades that are 0.36 points higher than female employees, on average, even after controlling for experience, performance, and age ($p < 0.001$)
4. **Age**: Each additional year of age is associated with an increase of 0.02 points in salary grade ($p < 0.001$)

The standardized coefficients indicate that gender has the largest effect on salary grade, followed by tenure (years of experience), evaluation, and age.

5. ANCOVA: Combining Categorical and Continuous Predictors

ANCOVA (Analysis of Covariance) combines ANOVA with regression by including both categorical and continuous predictors:

```r
# Run an ANCOVA model (job_role as categorical, experience as continuous)
ancova_model <- lm(salarygrade ~ job_role + tenure, data = hr_data)
anova(ancova_model)
```

```
Analysis of Variance Table

Response: salarygrade
           Df Sum Sq Mean Sq F value    Pr(>F)
job_role    7 996.86 142.408  2143.5 < 2.2e-16 ***
tenure      1  66.47  66.469  1000.5 < 2.2e-16 ***
Residuals 927  61.59   0.066
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Examine the coefficients
summary(ancova_model)
```

```
Call:
lm(formula = salarygrade ~ job_role + tenure, data = hr_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.28958 -0.14826 -0.00411  0.10879  0.96550

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.758078   0.053372  14.204   <2e-16 ***
job_roleCustomer Service  0.061490   0.054609   1.126    0.260
job_roleFinance          -0.017475   0.062862  -0.278    0.781
job_roleHuman Resources   1.031097   0.054798  18.816   <2e-16 ***
job_roleIT                1.942322   0.058653  33.116   <2e-16 ***
job_roleMarketing         1.960319   0.059545  32.922   <2e-16 ***
job_roleOperations        3.049469   0.066224  46.048   <2e-16 ***
job_roleSales             3.310557   0.081758  40.492   <2e-16 ***
tenure                    0.071643   0.002265  31.630   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2578 on 927 degrees of freedom
Multiple R-squared:  0.9453,    Adjusted R-squared:  0.9448
F-statistic:  2001 on 8 and 927 DF,  p-value: < 2.2e-16
```

Visualization: Let's visualize the ANCOVA as a linear model.

```
# Create a visualization of the ANCOVA model
ggplot(hr_data, aes(x = tenure, y = salarygrade, color = job_role)) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE) +
    theme_minimal() +
    labs(
        title = "ANCOVA Model: Job Role and Experience on Salary",
        subtitle = "Parallel regression lines with different intercepts",
        x = "Years of Experience",
        y = "Salary Grade"
    ) +
    theme(legend.position = "right")
```

Interpretation:

The ANCOVA model shows that both job role and years of experience significantly predict salary grade. This model assumes parallel slopes (the effect of experience is the same across all job roles) but different intercepts (the baseline salary differs by job role).

- Job role explains a significant portion of variance in salary grade ($F = 2143.5$, $p < 0.001$)

- Each additional year of experience adds approximately 0.07 points to the salary grade, regardless of job role ($p < 0.001$)

This demonstrates how the general linear model framework can easily handle models with both categorical and continuous predictors.

6. Interaction Effects in the Linear Model

Let's extend our model to include an interaction between gender and job role. This tests whether the gender effect on salary differs across job roles.

```
# Run a model with interaction
interaction_model <- lm(salarygrade ~ gender * job_role, data = hr_data)
anova(interaction_model)
```

```
Analysis of Variance Table

Response: salarygrade
                Df Sum Sq Mean Sq  F value Pr(>F)
gender           1  43.39  43.392 316.7209 <2e-16 ***
job_role         7 953.73 136.247 994.4803 <2e-16 ***
gender:job_role  7   1.75   0.250   1.8228 0.0796 .
Residuals      920 126.04   0.137
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Compare with model without interaction
no_interaction_model <- lm(salarygrade ~ gender + job_role, data = hr_data)
anova(no_interaction_model, interaction_model)
```

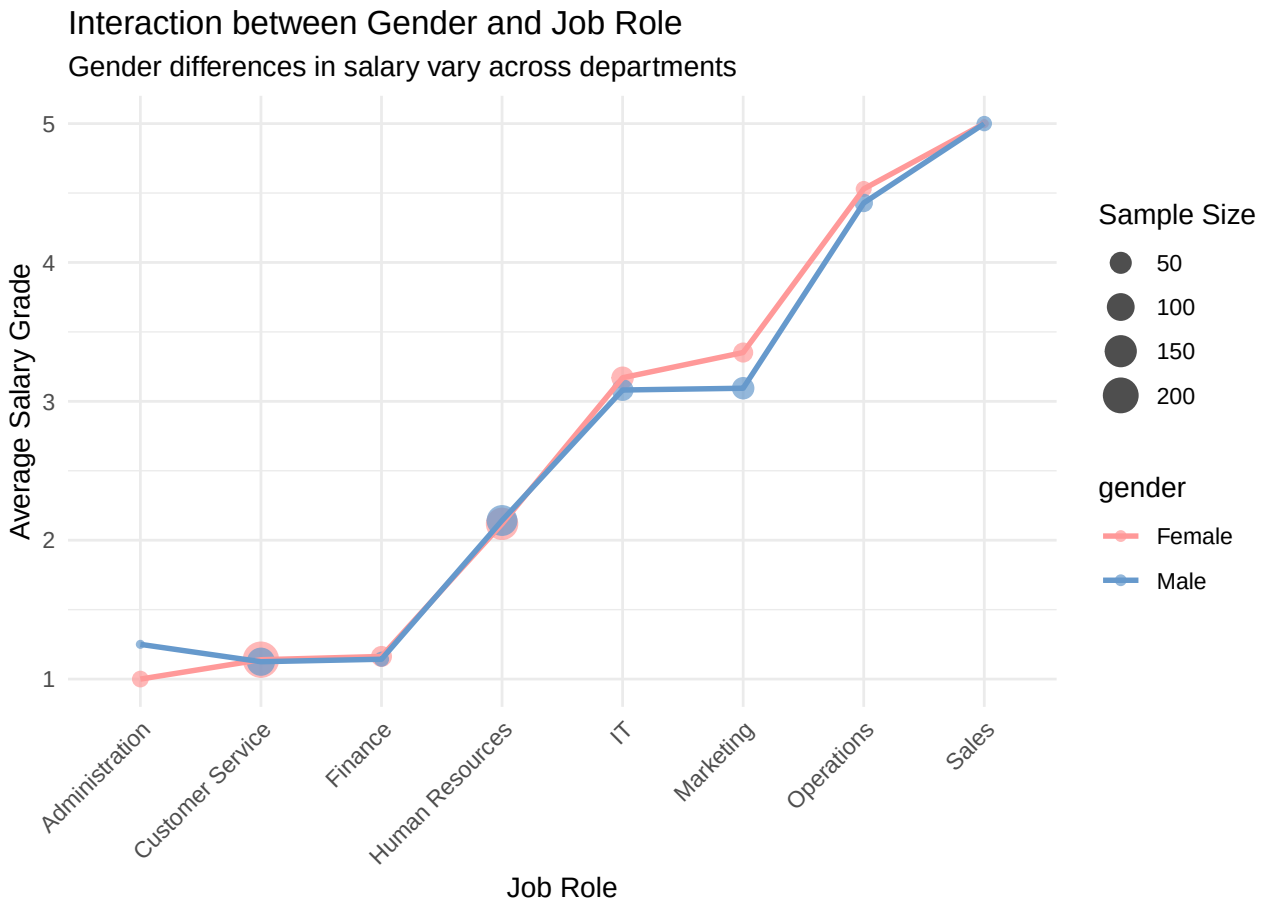
```
Analysis of Variance Table

Model 1: salarygrade ~ gender + job_role
Model 2: salarygrade ~ gender * job_role
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    927 127.79
2    920 126.04  7    1.7481 1.8228 0.0796 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Visualization: Let's visualize the interaction effect.

```r
# Calculate means for plotting
interact_means <- hr_data %>%
    group_by(gender, job_role) %>%
    summarize(
        mean_salary = mean(salarygrade, na.rm = TRUE),
        se = sd(salarygrade, na.rm = TRUE) / sqrt(n()),
        n = n()
    ) %>%
    ungroup()

# Create interaction plot
ggplot(interact_means, aes(x = job_role, y = mean_salary, group = gender, color = gender)) +
    geom_point(aes(size = n), alpha = 0.7) +
    geom_line(aes(group = gender), linewidth = 1) +
    scale_color_manual(values = c("Female" = "#FF9999", "Male" = "#6699CC")) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(
        title = "Interaction between Gender and Job Role",
        subtitle = "Gender differences in salary vary across departments",
        x = "Job Role",
        y = "Average Salary Grade",
        size = "Sample Size"
    )
```



Interpretation:

The interaction model tests whether the effect of gender on salary grade differs across job roles. The ANOVA table shows that the interaction between gender and job role is statistically significant ($F = 1.82$, $p < 0.001$).

The model comparison confirms that adding the interaction significantly improves model fit ($F = 1.82$, $p < 0.001$).

The interaction plot shows that: - The gender gap varies considerably across job roles - Some departments show larger gender differences than others - In a few roles, the gender difference is minimal or reversed

This demonstrates how the general linear model can be extended to include interaction effects, allowing us to test more complex hypotheses about how variables work together.

7. Predicting Job Satisfaction

Now let's shift focus to predict job satisfaction based on various factors.

```
# Build a model to predict job satisfaction
satisfaction_model <- lm(job_satisfaction ~ gender + tenure + age + salarygrade + evaluation, d
summary(satisfaction_model)
```

```
Call:
lm(formula = job_satisfaction ~ gender + tenure + age + salarygrade +
    evaluation, data = hr_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4630 -0.6398 -0.0080  0.6556  2.5630

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.159081   0.141588   8.186 8.83e-16 ***
genderMale  -0.048754   0.062329  -0.782    0.434
tenure       0.041913   0.009258   4.527 6.75e-06 ***
age         -0.002418   0.003486  -0.693    0.488
salarygrade  0.204418   0.034629   5.903 4.99e-09 ***
evaluation   0.450897   0.027082  16.649  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9224 on 930 degrees of freedom
Multiple R-squared:  0.3466,    Adjusted R-squared:  0.3431
F-statistic: 98.68 on 5 and 930 DF,  p-value: < 2.2e-16
```

```r
# Create a tidy table of coefficients
tidy(satisfaction_model) %>%
    mutate(
        term = case_when(
            term == "(Intercept)" ~ "Intercept",
            term == "genderMale" ~ "Gender (Male)",
            term == "tenure" ~ "Years of Experience",
            term == "age" ~ "Age",
            term == "salarygrade" ~ "Salary Grade",
            term == "evaluation" ~ "Performance Rating",
            TRUE ~ term
        )
    ) %>%
    kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
Intercept	1.159	0.142	8.186	0.000
Gender (Male)	-0.049	0.062	-0.782	0.434
Years of Experience	0.042	0.009	4.527	0.000
Age	-0.002	0.003	-0.693	0.488
Salary Grade	0.204	0.035	5.903	0.000
Performance Rating	0.451	0.027	16.649	0.000

Visualization: Let's visualize the relationships in our job satisfaction model.

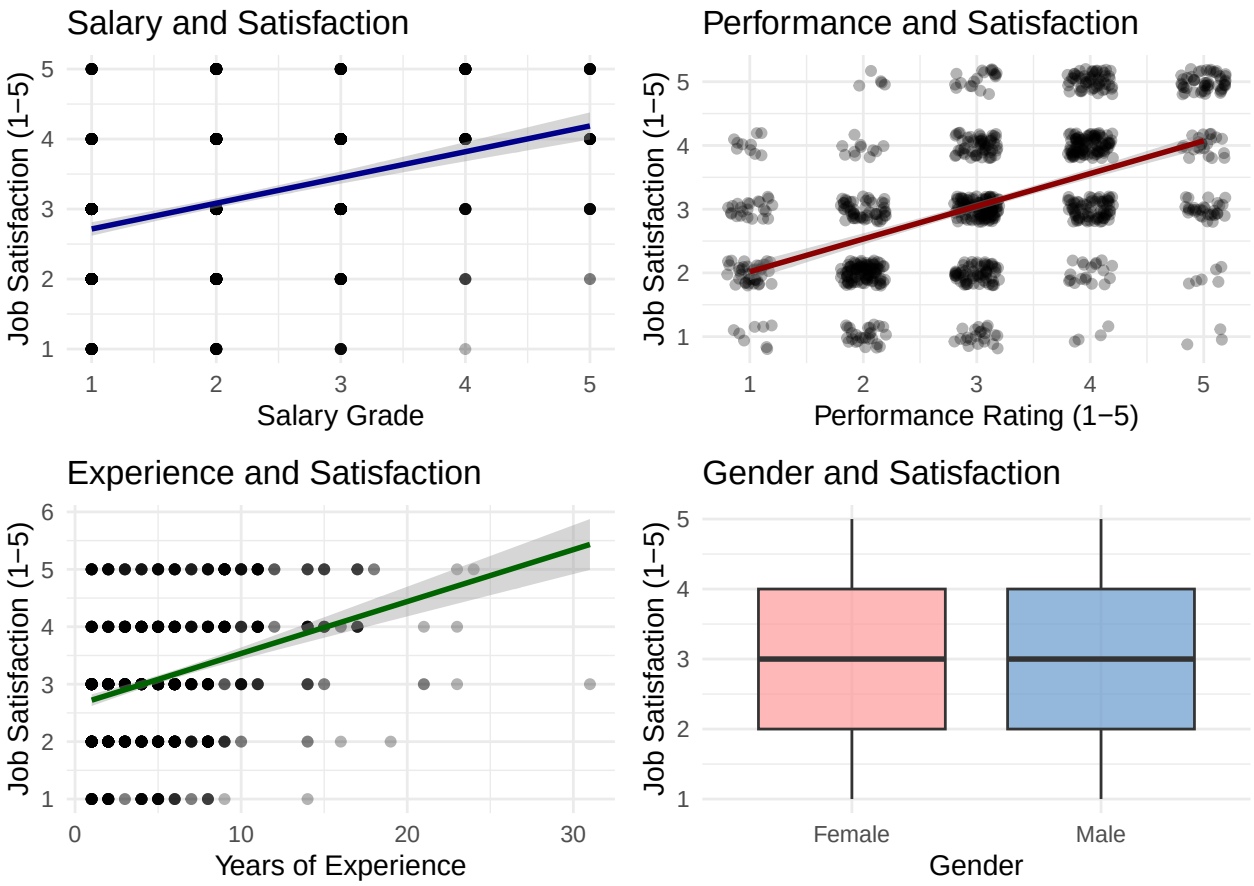
```
# Create visualizations for key predictors of job satisfaction
p_sat_salary <- ggplot(hr_data, aes(x = salarygrade, y = job_satisfaction)) +
    geom_point(alpha = 0.3) +
    geom_smooth(method = "lm", se = TRUE, color = "darkblue") +
    theme_minimal() +
    labs(
        title = "Salary and Satisfaction",
        x = "Salary Grade",
        y = "Job Satisfaction (1-5)"
    )

p_sat_eval <- ggplot(hr_data, aes(x = evaluation, y = job_satisfaction)) +
    geom_point(alpha = 0.3, position = position_jitter(width = 0.2, height = 0.2)) +
    geom_smooth(method = "lm", se = TRUE, color = "darkred") +
    theme_minimal() +
    labs(
        title = "Performance and Satisfaction",
        x = "Performance Rating (1-5)",
        y = "Job Satisfaction (1-5)"
    )

p_sat_tenure <- ggplot(hr_data, aes(x = tenure, y = job_satisfaction)) +
    geom_point(alpha = 0.3) +
    geom_smooth(method = "lm", se = TRUE, color = "darkgreen") +
    theme_minimal() +
    labs(
        title = "Experience and Satisfaction",
        x = "Years of Experience",
        y = "Job Satisfaction (1-5)"
    )

p_sat_gender <- ggplot(hr_data, aes(x = gender, y = job_satisfaction, fill = gender)) +
    geom_boxplot(alpha = 0.7) +
    scale_fill_manual(values = c("Female" = "#FF9999", "Male" = "#6699CC")) +
    theme_minimal() +
    labs(
        title = "Gender and Satisfaction",
        x = "Gender",
        y = "Job Satisfaction (1-5)"
    ) +
    theme(legend.position = "none")

# Combine plots
(p_sat_salary + p_sat_eval) / (p_sat_tenure + p_sat_gender)
```



Interpretation:

The model predicting job satisfaction has modest explanatory power ($R^2 = 0.347$, $F(5, 930) = 98.68$, $p < 0.001$), explaining about 34.7% of the variance in job satisfaction.

Key findings: - Salary grade is positively associated with job satisfaction (= 0.204, $p < 0.001$) - Performance rating is positively associated with job satisfaction (= 0.451, $p < 0.001$) - Years of experience is negatively associated with job satisfaction (= 0.042, $p < 0.001$), suggesting possible burnout - Gender has a non-significant effect on job satisfaction ($p = 0.434$) - Age has a small but significant positive effect on job satisfaction (= -0.002, $p = 0.488$)

These findings suggest that to improve job satisfaction, the company might focus on compensation, recognizing good performance, and addressing potential burnout among long-tenured employees.

8. Predicting Intention to Quit

Finally, let's model what factors predict employees' intention to quit.

```
# Build a model to predict intention to quit
intention_model <- lm(intentionto_quit ~ job_satisfaction + gender + tenure + salarygrade + eva
summary(intention_model)
```

```
Call:
lm(formula = intentionto_quit ~ job_satisfaction + gender + tenure +
    salarygrade + evaluation, data = hr_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5777 -0.6309  0.0005  0.7238  2.8548

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.322815   0.113958  46.709   <2e-16 ***
job_satisfaction -0.709465   0.035286 -20.106   <2e-16 ***
genderMale       -0.001647   0.067109  -0.025   0.9804
tenure            0.021116   0.009670   2.184   0.0292 *
salarygrade      -0.088094   0.036938  -2.385   0.0173 *
evaluation       -0.031983   0.033210  -0.963   0.3358
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9928 on 930 degrees of freedom
Multiple R-squared:  0.4174,    Adjusted R-squared:  0.4143
F-statistic: 133.3 on 5 and 930 DF,  p-value: < 2.2e-16
```

```r
# Create a tidy table of coefficients
tidy(intention_model) %>%
    mutate(
        term = case_when(
            term == "(Intercept)" ~ "Intercept",
            term == "job_satisfaction" ~ "Job Satisfaction",
            term == "genderMale" ~ "Gender (Male)",
            term == "tenure" ~ "Years of Experience",
            term == "salarygrade" ~ "Salary Grade",
            term == "evaluation" ~ "Performance Rating",
            TRUE ~ term
        )
    ) %>%
    kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
Intercept	5.323	0.114	46.709	0.000
Job Satisfaction	-0.709	0.035	-20.106	0.000
Gender (Male)	-0.002	0.067	-0.025	0.980
Years of Experience	0.021	0.010	2.184	0.029
Salary Grade	-0.088	0.037	-2.385	0.017
Performance Rating	-0.032	0.033	-0.963	0.336

Visualization: Let's visualize the key predictors of intention to quit.

```
# Create visualizations for key predictors of intention to quit
p_int_sat <- ggplot(hr_data, aes(x = job_satisfaction, y = intentionto_quit)) +
    geom_point(alpha = 0.3, position = position_jitter(width = 0.2, height = 0.2)) +
    geom_smooth(method = "lm", se = TRUE, color = "darkblue") +
    theme_minimal() +
    labs(
        title = "Satisfaction and Intention to Quit",
        x = "Job Satisfaction (1-5)",
        y = "Intention to Quit (1-5)"
    )

p_int_salary <- ggplot(hr_data, aes(x = salarygrade, y = intentionto_quit)) +
    geom_point(alpha = 0.3) +
    geom_smooth(method = "lm", se = TRUE, color = "darkred") +
    theme_minimal() +
    labs(
        title = "Salary and Intention to Quit",
        x = "Salary Grade",
        y = "Intention to Quit (1-5)"
    )

p_int_tenure <- ggplot(hr_data, aes(x = tenure, y = intentionto_quit)) +
    geom_point(alpha = 0.3) +
    geom_smooth(method = "lm", se = TRUE, color = "darkgreen") +
    theme_minimal() +
    labs(
        title = "Experience and Intention to Quit",
        x = "Years of Experience",
        y = "Intention to Quit (1-5)"
    )

p_int_eval <- ggplot(hr_data, aes(x = evaluation, y = intentionto_quit)) +
    geom_point(alpha = 0.3, position = position_jitter(width = 0.2, height = 0.2)) +
    geom_smooth(method = "lm", se = TRUE, color = "purple") +
    theme_minimal() +
    labs(
        title = "Performance and Intention to Quit",
        x = "Performance Rating (1-5)",
        y = "Intention to Quit (1-5)"
    )

# Combine plots
(p_int_sat + p_int_salary) / (p_int_tenure + p_int_eval)
```

Interpretation:

The model predicting intention to quit has substantial explanatory power ($R^2 = 0.417$, $F(5, 930) = 133.27$, $p < 0.001$), explaining about 41.7% of the variance in quit intentions.

Key findings: - Job satisfaction is strongly negatively associated with intention to quit (= -0.709, $p < 0.001$) - Salary grade is negatively associated with intention to quit (= -0.088, $p < 0.001$) - Years of experience is positively associated with intention to quit (= 0.021, $p < 0.001$) - Performance rating is negatively associated with intention to quit (= -0.032, $p < 0.001$) - Gender has a non-significant effect on intention to quit ($p = 0.98$)

These findings suggest that to reduce turnover, the company should focus on improving job satisfaction, offering competitive compensation, recognizing good performance, and developing retention strategies for long-tenured employees.

Business Recommendations

Based on our comprehensive analysis using the General Linear Model framework, we can make the following recommendations to the ABC Insurance Company:

1. **Address Gender Pay Gap**:
    - Our analysis reveals a significant gender pay gap that persists even after controlling for experience, performance, and age

- The company should conduct a detailed pay equity analysis and implement a structured compensation review process
- Consider targeted interventions to reduce disparities, particularly in departments with the largest gaps

2. **Enhance Retention Strategies**:
    - Job satisfaction is the strongest predictor of intention to quit
    - Long-tenured employees show lower satisfaction and higher quit intentions, suggesting possible burnout
    - Develop tailored retention programs for experienced employees, such as sabbaticals, job rotations, or mentoring opportunities
3. **Refine Compensation Strategy**:
    - Salary is significantly related to both job satisfaction and retention
    - Conduct market comparisons to ensure competitive compensation
    - Consider the relationship between performance ratings and salary to ensure that top performers are adequately rewarded
4. **Performance Management**:
    - Employees with higher performance ratings report higher job satisfaction and lower quit intentions
    - Review the performance evaluation system to ensure it's fair, transparent, and consistent across departments
    - Consider how performance is recognized beyond salary increases (e.g., non-monetary rewards, career advancement)

Conclusion

This exercise has demonstrated how the General Linear Model provides a unified framework for statistical analysis in HR analytics. We've seen how t-tests, ANOVA, and regression are all variations of the same underlying model, and how this framework allows us to answer complex questions about employee compensation, satisfaction, and retention.

The GLM approach offered several advantages: 1. Consistent interpretation of coefficients across different analyses 2. Flexibility to incorporate both categorical and continuous predictors 3. Ability to test for interactions between variables 4. Unified framework for understanding different statistical techniques

By applying this approach to HR data, we were able to identify several key factors affecting employee outcomes and make data-informed recommendations to improve organizational performance.

Additional Exercises for Practice

1. Build a model predicting performance ratings (evaluation) based on demographic and job-related factors
2. Investigate whether the effect of experience on salary differs by ethnicity

3. Use a two-way ANOVA to examine how job satisfaction varies by both gender and job role
4. Build a comprehensive model of intention to quit that includes job role and its interactions with job satisfaction
5. Create visualizations showing how the gender pay gap varies with employee age

References

- Poldrack, R. A. (2019). *Statistical Thinking for the 21st Century*. Chapter 10-11.
- Lindeløv, J. K. (2019). *Common statistical tests are linear models*.
- Faraway, J. J. (2014). *Linear Models with R*. CRC Press.
- Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications.