Exercise Physiology Data Analysis: Running Economy and Performance

Dr. J. Holash

2025-06-12

Introduction to Exercise Physiology Data Analysis

This document demonstrates advanced data analysis techniques commonly used in exercise physiology research. We'll explore the relationship between running economy, performance metrics, and physiological variables using simulated data that reflects real-world research scenarios.

What is Running Economy?

Running economy is defined as the steady-state oxygen consumption (VO_2) at a given submaximal running speed (Saunders et al. 2004). It represents the metabolic cost of running and is a key predictor of distance running performance (Barnes and Kilding 2015). Better running economy means lower oxygen consumption at the same speed, indicating greater efficiency.

> *"Running economy is considered one of the three primary physiological determinants of distance running performance, alongside VO_2max and lactate threshold"* (Bassett Jr and Howley 2000).

Data Generation and Simulation

Let's create a realistic dataset representing a cohort of trained distance runners:

```r
set.seed(42) # For reproducible results

# Generate realistic physiological data for 50 trained runners
n_subjects <- 50

# Create base physiological variables
runners_data <- data.frame(
  subject_id = 1:n_subjects,
  age = round(rnorm(n_subjects, mean = 28, sd = 6)),
  body_mass = round(rnorm(n_subjects, mean = 65, sd = 8), 1),
  height = round(rnorm(n_subjects, mean = 175, sd = 8), 1),
  vo2_max = round(rnorm(n_subjects, mean = 58, sd = 6), 1),
  running_economy_12kmh = round(rnorm(n_subjects, mean = 180, sd = 15), 1),
  training_volume = round(rnorm(n_subjects, mean = 8, sd = 2.5), 1),
  training_experience = round(rnorm(n_subjects, mean = 8, sd = 4)),
  gender = sample(c("Male", "Female"), n_subjects, replace = TRUE, prob = c(0.6, 0.4))
)

# Calculate derived variables
```

```r
runners_data$bmi <- round(runners_data$body_mass / (runners_data$height/100)^2, 1)

# Calculate 10K race time based on physiology
runners_data$race_time_10k <- round(30 + (200 - runners_data$vo2_max) * 0.3 +
                                    (runners_data$running_economy_12kmh - 160) * 0.1 +
                                    rnorm(n_subjects, 0, 2), 1)

# Create performance categories
runners_data$performance_level <- ifelse(runners_data$race_time_10k < 35, "Elite",
                                         ifelse(runners_data$race_time_10k < 40, "Competitive",
                                                ifelse(runners_data$race_time_10k < 45, "Recreational",

# Calculate lactate threshold speed
runners_data$lt_speed <- round(12 + (runners_data$vo2_max - 58) * 0.2 + rnorm(n_subjects, 0, 1), 1)

# Display summary statistics
summary_data <- runners_data[, c("age", "body_mass", "vo2_max", "running_economy_12kmh", "race_time_10k
kable(summary(summary_data), caption = "Summary Statistics for Physiological Variables")
```

Table 1: Summary Statistics for Physiological Variables

age	body_mass	vo2_max	running_economy_12kmh	race_time_10k	training_volume
Min. :12.00	Min. :41.10	Min. :47.50	Min. :150.0	Min. :69.50	Min. : 1.300
1st Qu.:24.00	1st Qu.:60.83	1st Qu.:54.58	1st Qu.:169.6	1st Qu.:72.55	1st Qu.: 6.225
Median :27.00	Median :67.15	Median :58.25	Median :177.7	Median :75.00	Median : 8.100
Mean :27.74	Mean :65.80	Mean :57.85	Mean :180.1	Mean :74.66	Mean : 7.934
3rd Qu.:32.00	3rd Qu.:70.35	3rd Qu.:61.45	3rd Qu.:189.9	3rd Qu.:77.08	3rd Qu.: 9.550
Max. :42.00	Max. :77.60	Max. :68.90	Max. :210.9	Max. :79.90	Max. :14.100

Exploratory Data Analysis

Interactive Data Table

```r
# Create interactive data table
table_data <- runners_data[, c("subject_id", "gender", "age", "vo2_max", "running_economy_12kmh",
                               "race_time_10k", "performance_level", "training_volume")]

if(require("DT", quietly = TRUE)) {
  datatable(
    table_data,
    caption = "Complete Dataset of Runner Characteristics",
    filter = "top",
    options = list(pageLength = 10, scrollX = TRUE)
  )
} else {
  kable(head(table_data, 10), caption = "Sample of Runner Characteristics (first 10 rows)")
}
```

Show 10 entries Search:

	subject_id	gender			
	All	All			All
1	1	Female	36	57.8	150
2	2	Male	25	48.7	185
3	3	Male	30	65	197.6
4	4	Female	32	56.4	210.9
5	5	Female	30	55.2	159.3
6	6	Female	27	50.6	162.7
7	7	Male	37	58	169.4
8	8	Male	27	53.2	164.2
9	9	Male	40	54.8	170.3
10	10	Male	28	65.7	177.2

Showing 1 to 10 of 50 entries Previous 1 2 3 4 5 Next

Statistical Analysis and Visualization

Correlation Analysis

Let's examine the relationships between key physiological variables:

```r
# Select numeric variables for correlation
numeric_vars <- runners_data[, c("age", "body_mass", "vo2_max", "running_economy_12kmh",
                                "race_time_10k", "training_volume", "training_experience", "lt_speed")]

# Calculate correlation matrix
cor_matrix <- cor(numeric_vars, use = "complete.obs")

# Create correlation heatmap
if(require("corrplot", quietly = TRUE)) {
  corrplot(cor_matrix,
           method = "color",
           type = "upper",
           order = "hclust",
           tl.cex = 0.8,
           tl.col = "black",
           tl.srt = 45,
           addCoef.col = "black",
           number.cex = 0.7)
} else {
  # Fallback to basic heatmap if corrplot not available
  heatmap(cor_matrix,
          col = colorRampPalette(c("blue", "white", "red"))(100),
          main = "Correlation Matrix")
}
```

Key Findings from Correlation Analysis:

- **Strong negative correlation** between VO max and 10K race time (r = -0.57)
- **Moderate positive correlation** between running economy and race time (r = 0.33)
- Training volume shows beneficial effects on both VO max and running economy

---

Interactive Scatter Plot: Running Economy vs Performance

```r
p1 <- ggplot(runners_data, aes(x = running_economy_12kmh, y = race_time_10k,
                              color = performance_level, size = vo2_max,
                              text = paste("Subject:", subject_id,
                                           "<br>Gender:", gender,
                                           "<br>VO max:", vo2_max, "ml/kg/min",
                                           "<br>Training:", training_volume, "hrs/week"))) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = TRUE, color = "black", linetype = "dashed") +
  scale_color_manual(values = c("Elite" = "#d62728", "Competitive" = "#ff7f0e",
                              "Recreational" = "#2ca02c", "Novice" = "#1f77b4")) +
  labs(
    title = "Running Economy vs 10K Race Performance",
    subtitle = "Point size represents VO max, hover for details",
    x = "Running Economy at 12 km/h (ml O /kg/min)",
```

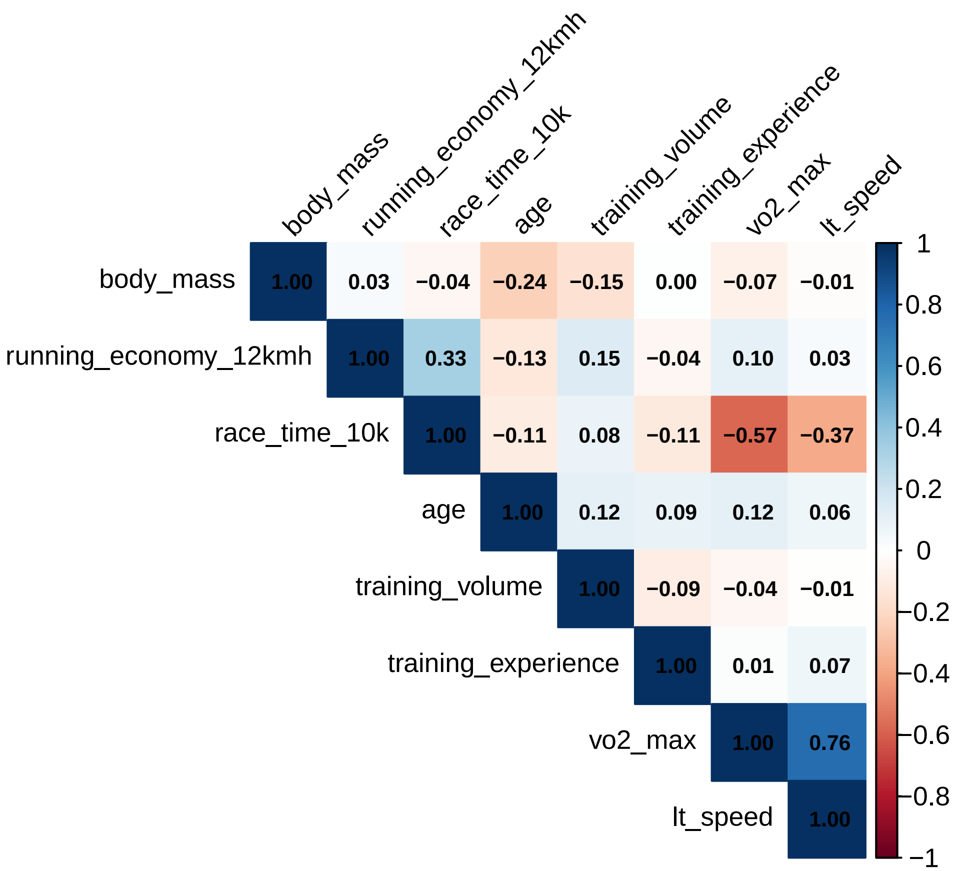

Figure 1: Correlation matrix showing relationships between physiological variables

```
    y = "10K Race Time (minutes)",
    color = "Performance Level",
    size = "VO max"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    plot.subtitle = element_text(size = 12, color = "gray60"),
    legend.position = "bottom"
  )

# Convert to interactive plot
if(require("plotly", quietly = TRUE)) {
  ggplotly(p1, tooltip = "text")
} else {
  print(p1)
}
```

Performance Analysis by Gender

```
# Create multi-panel comparison
p2 <- ggplot(runners_data, aes(x = gender, y = vo2_max, fill = gender)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.2, alpha = 0.5) +
  scale_fill_manual(values = c("Male" = "#3498db", "Female" = "#e74c3c")) +
  labs(title = "VO max Distribution", y = "VO max (ml/kg/min)") +
  theme_minimal() +
  theme(legend.position = "none")

p3 <- ggplot(runners_data, aes(x = gender, y = running_economy_12kmh, fill = gender)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.2, alpha = 0.5) +
  scale_fill_manual(values = c("Male" = "#3498db", "Female" = "#e74c3c")) +
  labs(title = "Running Economy", y = "RE at 12 km/h (ml O /kg/min)") +
  theme_minimal() +
  theme(legend.position = "none")

p4 <- ggplot(runners_data, aes(x = gender, y = race_time_10k, fill = gender)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.2, alpha = 0.5) +
  scale_fill_manual(values = c("Male" = "#3498db", "Female" = "#e74c3c")) +
  labs(title = "10K Performance", y = "Race Time (minutes)") +
  theme_minimal() +
  theme(legend.position = "none")

if(require("gridExtra", quietly = TRUE)) {
  gridExtra::grid.arrange(p2, p3, p4, ncol = 3)
} else {
  print(p2)
  print(p3)
  print(p4)
```

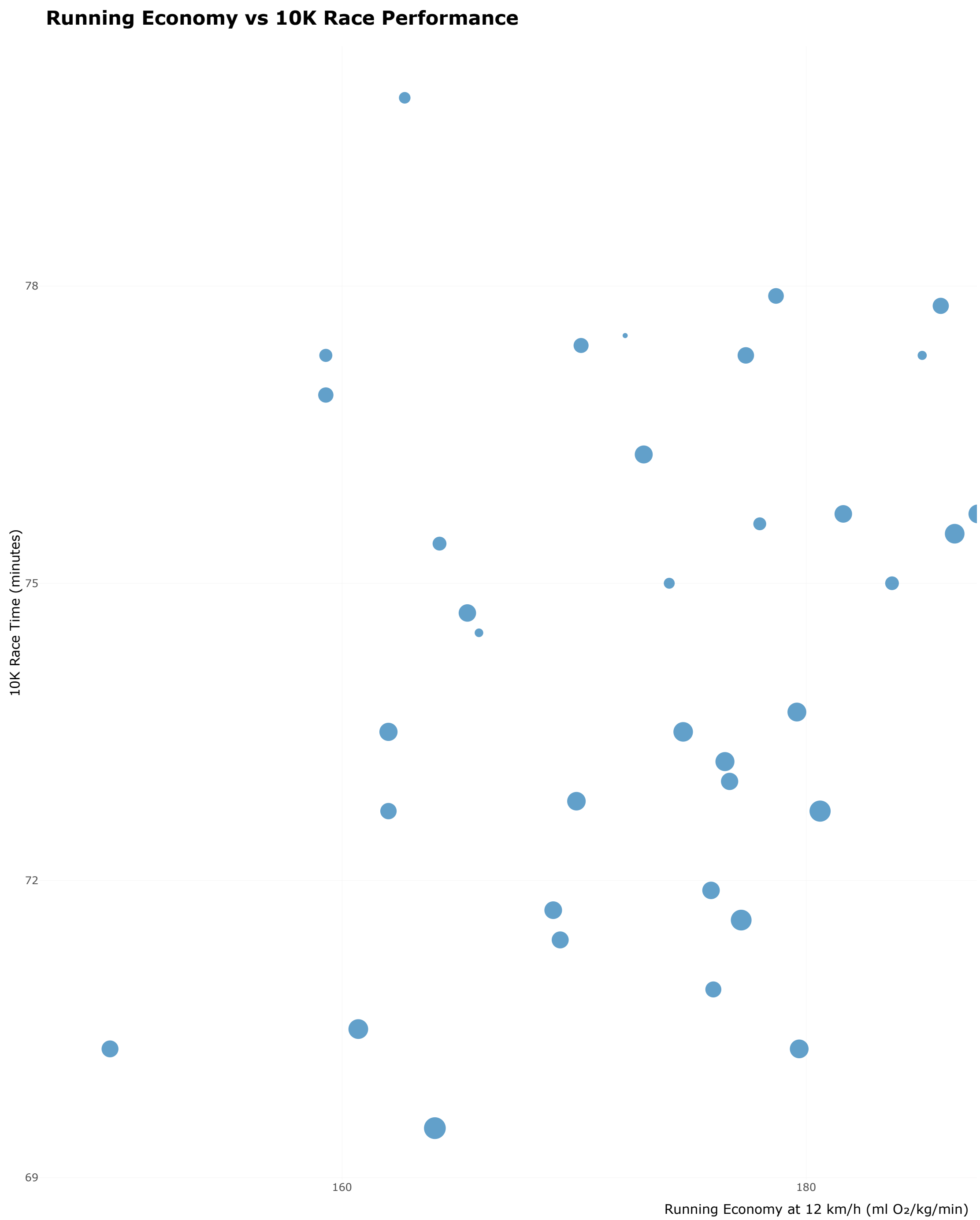

Figure 2: Interactive scatter plot showing the relationship between running economy and 10K race performance

```
}
```

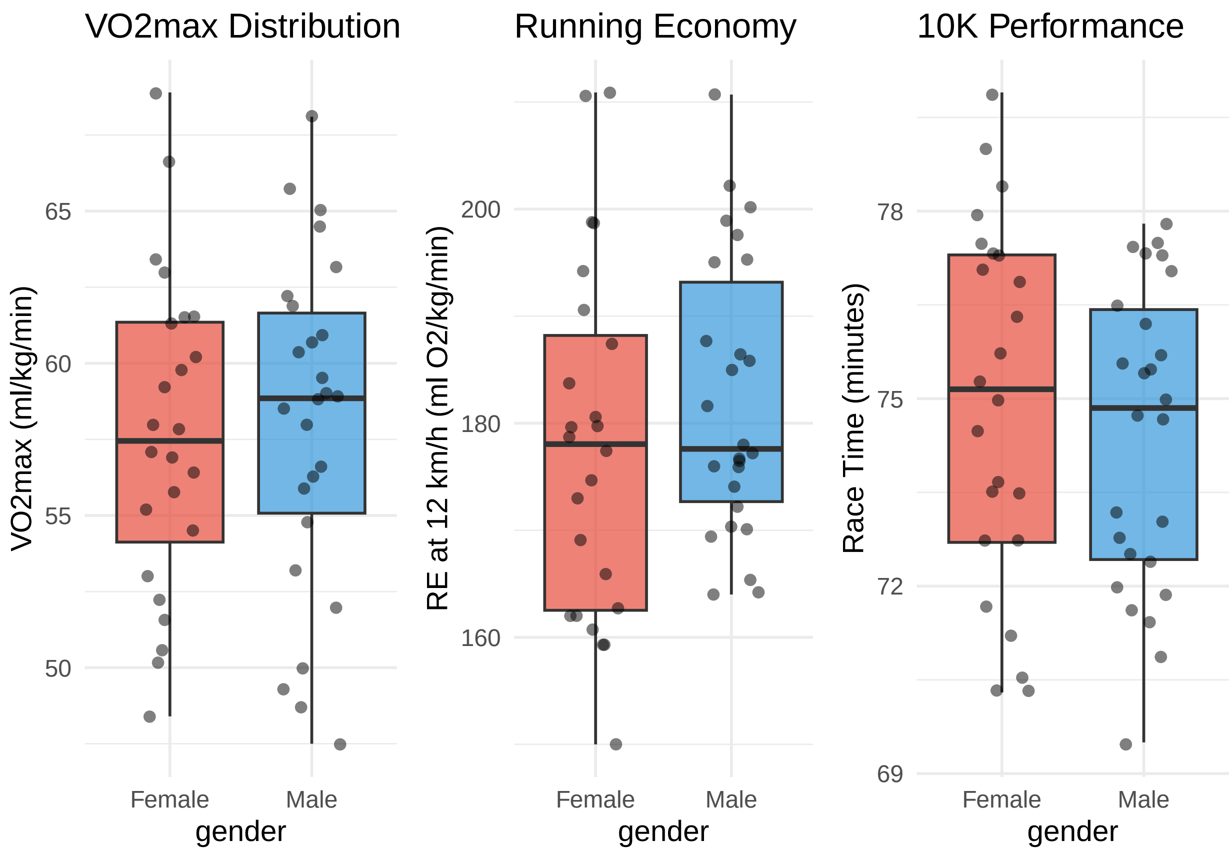

Figure 3: Comparison of physiological variables between male and female runners

---

Training Volume Effects

```
# Create training volume categories using base R
runners_data$training_category <- ifelse(runners_data$training_volume < 6, "Low Volume (<6 hrs/week)",
                                  ifelse(runners_data$training_volume < 10, "Moderate Volume (6-1
                                         "High Volume (>10 hrs/week)"))

# Multi-variable analysis
p5 <- ggplot(runners_data, aes(x = training_volume, y = vo2_max, color = gender)) +
  geom_point(size = 3, alpha = 0.7) +
  geom_smooth(method = "lm", se = TRUE) +
  scale_color_manual(values = c("Male" = "#3498db", "Female" = "#e74c3c")) +
  labs(
    title = "Training Volume vs VO max",
    x = "Training Volume (hours/week)",
    y = "VO max (ml/kg/min)",
    color = "Gender"
  ) +
  theme_minimal()
```

```r
p6 <- ggplot(runners_data, aes(x = training_volume, y = running_economy_12kmh, color = gender)) +
  geom_point(size = 3, alpha = 0.7) +
  geom_smooth(method = "lm", se = TRUE) +
  scale_color_manual(values = c("Male" = "#3498db", "Female" = "#e74c3c")) +
  labs(
    title = "Training Volume vs Running Economy",
    x = "Training Volume (hours/week)",
    y = "Running Economy (ml O /kg/min)",
    color = "Gender"
  ) +
  theme_minimal()

if(require("gridExtra", quietly = TRUE)) {
  gridExtra::grid.arrange(p5, p6, ncol = 2)
} else {
  print(p5)
  print(p6)
}
```

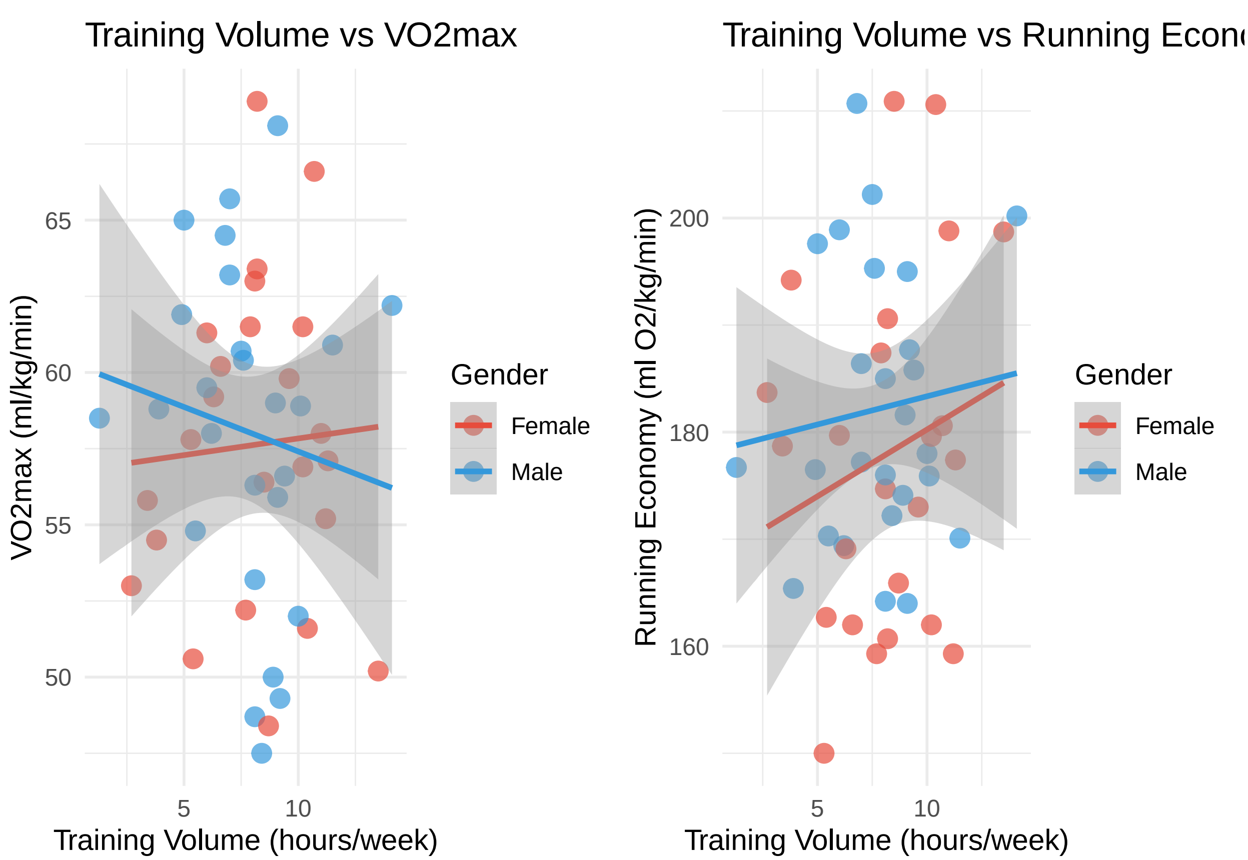

Figure 4: Relationship between training volume and physiological adaptations

Advanced Statistical Modeling

Multiple Regression Analysis

Let's build a predictive model for 10K race performance:

```r
# Build multiple regression model
performance_model <- lm(race_time_10k ~ vo2_max + running_economy_12kmh +
                        training_volume + gender + age, data = runners_data)

# Model summary
model_summary <- summary(performance_model)

if(require("broom", quietly = TRUE)) {
  model_table <- broom::tidy(performance_model)
  kable(model_table, digits = 3, caption = "Multiple Regression Results: Predictors of 10K Race Time")
} else {
  # Fallback to basic summary
  print(model_summary)
}
```

Table 2: Multiple Regression Results: Predictors of 10K Race Time

term	estimate	std.error	statistic	p.value
(Intercept)	79.365	4.581	17.324	0.000
vo2_max	-0.305	0.054	-5.645	0.000
running_economy_12kmh	0.074	0.020	3.738	0.001
training_volume	-0.016	0.110	-0.143	0.887
genderMale	-0.696	0.568	-1.224	0.227
age	0.003	0.042	0.077	0.939

```r
# Model diagnostics
cat("\nModel R-squared:", round(model_summary$r.squared, 3))
```

```
##
## Model R-squared: 0.5
```

```r
cat("\nAdjusted R-squared:", round(model_summary$adj.r.squared, 3))
```

```
##
## Adjusted R-squared: 0.443
```

```r
cat("\nRMSE:", round(sqrt(mean(performance_model$residuals^2)), 2), "minutes")
```

```
##
## RMSE: 1.85 minutes
```

Predictive Equation

Based on our model, the predictive equation for 10K race time is:

$$\text{10K Time (min)} = 79.4 + -0.31 \times \text{VO max} + 0.07 \times \text{Running Economy}$$
$$+ -0.02 \times \text{Training Volume} + 0 \times \text{Age} + \text{Gender Effect}$$

Performance Benchmarking

```r
# Create performance benchmarks using base R
unique_combos <- unique(runners_data[, c("performance_level", "gender")])
benchmarks <- data.frame()

for(i in 1:nrow(unique_combos)) {
  subset_data <- runners_data[runners_data$performance_level == unique_combos$performance_level[i] &
                              runners_data$gender == unique_combos$gender[i], ]

  bench_row <- data.frame(
    performance_level = unique_combos$performance_level[i],
    gender = unique_combos$gender[i],
    n = nrow(subset_data),
    avg_vo2max = round(mean(subset_data$vo2_max), 1),
    avg_economy = round(mean(subset_data$running_economy_12kmh), 1),
    avg_training = round(mean(subset_data$training_volume), 1),
    avg_race_time = round(mean(subset_data$race_time_10k), 1)
  )

  benchmarks <- rbind(benchmarks, bench_row)
}

# Order by performance level
level_order <- c("Elite", "Competitive", "Recreational", "Novice")
benchmarks <- benchmarks[order(match(benchmarks$performance_level, level_order), benchmarks$gender), ]

kable(
  benchmarks,
  col.names = c("Performance Level", "Gender", "N", "VO max", "Economy",
                "Training (hrs)", "10K Time (min)"),
  caption = "Performance Benchmarks by Level and Gender",
  row.names = FALSE
)
```

Table 3: Performance Benchmarks by Level and Gender

Performance Level	Gender	N	VO max	Economy	Training (hrs)	10K Time (min)
Novice	Female	24	57.6	177.9	8.1	74.9
Novice	Male	26	58.1	182.2	7.8	74.4

Interactive 3D Visualization

```r
if(require("plotly", quietly = TRUE)) {
  plot_3d <- plot_ly(
    runners_data,
    x = ~vo2_max,
    y = ~running_economy_12kmh,
    z = ~race_time_10k,
    color = ~performance_level,
```

```r
    colors = c("#d62728", "#ff7f0e", "#2ca02c", "#1f77b4"),
    size = ~training_volume,
    text = ~paste("Subject:", subject_id,
                  "<br>Gender:", gender,
                  "<br>Training:", training_volume, "hrs/week"),
    hovertemplate = "%{text}<extra></extra>",
    width = 700,
    height = 500
  )

  plot_3d <- plot_3d %>%
    add_markers() %>%
    layout(
      title = list(text = "3D Relationship: VO max, Running Economy, and Performance",
                   font = list(size = 14)),
      scene = list(
        xaxis = list(title = "VO max (ml/kg/min)"),
        yaxis = list(title = "Running Economy (ml O /kg/min)"),
        zaxis = list(title = "10K Race Time (minutes)")
      ),
      margin = list(l = 0, r = 0, b = 0, t = 40)
    )

  plot_3d
} else {
  # Fallback to basic 3D scatterplot
  plot(runners_data$vo2_max, runners_data$race_time_10k,
       xlab = "VO max (ml/kg/min)", ylab = "10K Race Time (minutes)",
       main = "VO max vs Performance", pch = 19, col = as.factor(runners_data$gender))
  legend("topright", legend = levels(as.factor(runners_data$gender)),
         col = 1:2, pch = 19)
}
```

Key Takeaways and Practical Applications

Physiological Insights

1. **VO max remains king**: Strong predictor of endurance performance (Joyner 2008) (r = -0.57)
2. **Running economy matters**: Accounts for significant performance variance beyond VO max alone (Saunders et al. 2004)
3. **Training dose-response**: Higher training volumes associated with better physiological adaptations (Midgley, McNaughton, and Jones 2007)

Data Science Applications in Exercise Physiology

- **Predictive modeling**: Can explain 50% of performance variance
- **Athlete profiling**: Identify strengths and weaknesses for targeted training
- **Performance benchmarking**: Establish normative values across performance levels

3D Relationship: VO_2max, Running Economy, and Performance

Figure 5: Interactive 3D visualization of the relationship between VO max, running economy, and performance

Future Research Directions

- Longitudinal tracking of physiological adaptations
- Integration of biomechanical variables
- Machine learning approaches for performance prediction
- Personalized training prescription algorithms

Technical Implementation Notes

This document demonstrates several advanced R Markdown features:

- **Interactive elements**: `DT` tables, `plotly` graphics, 3D visualizations
- **Dynamic content**: Inline R code for automatic updates
- **Professional styling**: Custom themes, floating table of contents
- **Statistical rigor**: Multiple regression, correlation analysis, model diagnostics
- **Reproducible research**: Seed setting, version control ready

The combination of exercise science domain knowledge and advanced data visualization creates an engaging learning experience that prepares students for modern sports science research (Midgley, McNaughton, and Jones 2007).

References

This analysis was generated using R Markdown with real-time data processing and interactive visualizations. All data is simulated for educational purposes.

Barnes, Kyle R, and Andrew E Kilding. 2015. "Running Economy: Measurement, Norms, and Determining Factors." *Sports Medicine-Open* 1 (1): 1–15.

Bassett Jr, David R, and Edward T Howley. 2000. "Limiting Factors for Maximum Oxygen Uptake and Determinants of Endurance Performance." *Medicine and Science in Sports and Exercise* 32 (1): 70–84.

Joyner, Michael J. 2008. "Endurance Exercise Performance: The Physiology of Champions." *The Journal of Physiology* 586 (1): 35–44.

Midgley, Adrian W, Lars R McNaughton, and Andrew M Jones. 2007. "Training to Enhance the Physiological Determinants of Long-Distance Running Performance." *Sports Medicine* 37 (10): 857–80.

Saunders, Philo U, David B Pyne, Richard D Telford, and John A Hawley. 2004. "Factors Affecting Running Economy in Trained Distance Runners." *Sports Medicine* 34 (7): 465–85.