Applied Statistical Modeling

Module: ANOVA, ANCOVA, Randomized Blocks, LASSO

Author: Charalampos (Babis) Giannakis
Affiliation: MSc Business Analytics – Computational Intelligence, VU Amsterdam
Version: v1.0 **Project ID:** SM-2025-01
Date: November 7, 2025
Contact: github.com/CharalamposGiannakis

We start from structure. Complexity follows.

Introduction

This report presents a series of statistical analyses conducted in R, using RMarkdown for full reproducibility. The goal is to investigate how different explanatory factors influence a set of response variables across several experimental settings. The methods used include one-way and two-way ANOVA, randomized block designs, ANCOVA, and LASSO regression.

For each task, the relevant model is specified, fitted, and evaluated using appropriate diagnostics. The results are interpreted in the context of the research questions, with emphasis on model assumptions, effect interpretation, and predictive performance. The document is organized according to the structure of the problem set, progressing from exploratory analysis to model comparison and prediction.

Section 1 (theoretical)

a) We observe moisture Y_{ijk} for method $i \in \{1,2\}$, type $j \in \{1,2,3\}$, replicate $k \in \{1,2\}$ The full two–way ANOVA with interaction is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$$

Identifiability (sum-to-zero) restrictions:
$\sum_{i=1}^{2} \alpha_i = 0, \quad \sum_{j=1}^{3} \beta_j = 0, \quad \sum_{i=1}^{2} (\alpha\beta)_{ij} = 0 \ \forall j, \quad \sum_{j=1}^{3} (\alpha\beta)_{ij} = 0 \ \forall i.$

With $a = 2$, $b = 3$, $k = 2$ we have $N = abk = 12$ observations and

$df_{\text{method}} = 1, \quad df_{\text{type}} = 2, \quad df_{\text{int}} = 2, \quad df_E = 6, \quad df_T = 11$

Error mean square: $MS_E = SS_E/df_E = 0.6620/6 = 0.1103$

Interaction: $SS_{\text{int}} = df_{\text{int}} \cdot MS_{\text{int}} = 2 \cdot 0.1513 = 0.3026$
$F_{\text{int}} = MS_{\text{int}}/MS_E = 0.1513/0.1103 = 1.3717$

```
p_int <- pf(1.3717, 2, 6, lower.tail = FALSE)
```

$Pr(> F)_{int} = 0.323$

Method (main effect): $MS_{\text{meth}} = SS_{\text{meth}}/df_{\text{meth}} = 0.1141,$
$F_{\text{meth}} = MS_{\text{meth}}/MS_E = 0.1141/0.1103 = 1.0344$

```
p_method <- pf(1.0344, df1=1, df2=6, lower.tail = FALSE)
```

$Pr(> F)_{meth} = 0.348$

Type (main effect): $MS_{\text{type}} = F_{\text{type}} \cdot MS_E = 117.3810 \cdot 0.110333 = 12.9510$
$SS_{\text{type}} = df_{\text{type}} \cdot MS_{\text{type}} = 2 \cdot 12.951037 = 25.9020$

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
method	1	0.1141	0.1141	1.0341	0.348

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	2	25.9020	12.9510	117.3810	1.548e-05
method:type	2	0.3026	0.1513	1.3713	0.323
Residuals	6	0.6620	0.1103	-	-

Conclusion: The model is a two-way ANOVA with interaction:
method $F = 1.03$, $p = 0.35$ (not significant)
type $F = 117.38$, $p < 0.001$ (highly significant)
interaction $F = 1.37$, $p = 0.32$ (not significant)
Moisture content depends strongly on cheese type, but not on method and with no interaction.
$\Longrightarrow$ The reduced model is additive with type only.

b) From the ANOVA model above and the Two-Way ANOVA table we can calculate the estimator $\hat{\sigma}^2 = SS_E/(n - IJ) = 0.6620/(12 - 2 * 3) = 0.1103$

The coefficient of determination R^2:
$SS_{total} = SS_{meth} + SS_{type} + SS_{int} + SS_E = 0.1141 + 25.9020 + 0.3026 + 0.6620 = 26.9807$

Then: $R^2 = 1 - SS_E/SS_{total} = 1 - 0.6620/26.9807 = 0.9754(97,54\%)$

The adjusted R^2: $R^2_{adj} = 1 - \frac{SS_E/df_E}{SS_{total}/df_{total}} = 1 - \frac{0.6620/6}{26.9807/11} = 1 - \frac{0.1103}{2.4527} = 0.9550(95.50\%)$

For the full model, $R^2 = 0.9754$(= 97.54% of variability explained). The penalty yields $R^2_{adj} = 0.9550(95.50\%)$. Since method and interaction are not significant, a reduced type-only model would typically have an equal or higher R^2_{adj} with fewer parameters.

The "Residual standard error: 0.3322 on 6 degrees of freedom" in the second output is just:
$\hat{\sigma} = \sqrt{MS_E} = \sqrt{SS_E/df_E} = \sqrt{0.6620/6} = \sqrt{0.1103} = 0.3322 = RSE$

Link between $\hat{\sigma}$, RSE, and MSE:

$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I), \quad df_E = N - p - 1, \quad \text{SSE} = \sum e_i^2, \quad MS_E = \text{SSE}/df_E, \quad \hat{\sigma}^2 = MS_E.$

Residual standard error: $\text{RSE} = \sqrt{\hat{\sigma}^2} = \sqrt{MS_E}$.

c) From the coefficient table, the intercept corresponds to method 1 and type 1, with estimated mean moisture level $\hat{\mu}_{1,1} = 38.905$. The coefficients for type 2 and type 3 are both negative and highly significant (type2: -3.33, $p < 0.001$; type3: -2.40, $p < 0.001$), indicating that cheeses of type 2 and type 3 have significantly lower moisture content than type 1. The coefficient for method (method2) is very small and not significant ($p \approx 0.82$), so the method does not affect the moisture level. The interaction terms are also not significant, implying no interaction between method and type.

Conclusion: Moisture content depends strongly on cheese type but not on method. The highest moisture is obtained for **type 1**, regardless of the method used.
d) We model the moisture as $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$

with method $i \in \{1, 2\}$, type $j \in \{1, 2, 3\}$, two replicates per cell.

Assumptions: errors are independent, homoscedastic, and normal.

From the output of summary(cheesefit):

$\hat{n}_{11} = 38.9050,\ \hat{n}_{21} = 38.9850,\ \hat{n}_{12} = 35.5750,\ \hat{n}_{22} = 35.5500,\ \hat{n}_{13} = 36.5100,\ \hat{n}_{23} = 35.8700.$

The six fitted cell means are:

$$\hat{n}_{11} = \hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_1 + \hat{\gamma}_{11}, \quad \hat{n}_{21} = \hat{\mu} + \hat{\alpha}_2 + \hat{\beta}_1 + \hat{\gamma}_{21}, \quad \hat{n}_{12} = \hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_2 + \hat{\gamma}_{12},$$
$$\hat{n}_{22} = \hat{\mu} + \hat{\alpha}_2 + \hat{\beta}_2 + \hat{\gamma}_{22}, \quad \hat{n}_{13} = \hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_3 + \hat{\gamma}_{13}, \quad \hat{n}_{23} = \hat{\mu} + \hat{\alpha}_2 + \hat{\beta}_3 + \hat{\gamma}_{23}.$$

Using the **sum-to-zero** parametrization:

Overall mean: $\hat{\mu} = 36.8992$
Method effects: $\hat{\alpha}_1 = +0.0975,\ \hat{\alpha}_2 = -0.0975$
Type effects: $\hat{\beta}_1 = +2.0458,\ \hat{\beta}_2 = -1.3367,\ \hat{\beta}_3 = -0.7092$

Interaction effects ($\hat{\gamma}_{ij}$):

	j = 1	j = 2	j = 3
i=1	-0.1375	-0.0850	+0.2225
i=2	+0.1375	+0.0850	-0.2225

e) Estimating $G(\sigma) = (\sigma^2 + 1)^2$

Let S_Ω = SSE from the full two-way ANOVA, with $\nu = 6$ residual *df*

Unbiased variance estimate: $\widehat{\sigma^2} = S_\Omega/\nu$

$$S_\Omega/\sigma^2 \sim \chi^2_\nu \implies \mathbb{E}[S_\Omega] = \nu\sigma^2 \implies \hat{\sigma}^2 = S_\Omega/\nu$$

$$\mathbb{E}[(\chi^2_\nu)^2] = \nu(\nu+2) \implies \mathbb{E}[S_\Omega^2] = \sigma^4\nu(\nu+2) \implies \hat{\sigma}^4 = S_\Omega^2/[\,\nu(\nu+2)\,]$$

Unbiased σ^4 estimate: $\widehat{\sigma^4} = S_\Omega^2/[\nu(\nu+2)]$

Since $G(\sigma) = \sigma^4 + 2\sigma^2 + 1$, we have:

Unbiased estimator: $\widehat{G}_{\text{unb}} = \frac{S_\Omega^2}{\nu(\nu+2)} + 2\,\frac{S_\Omega}{\nu} + 1$

Plug-in estimator: $\widehat{G}_{\text{plug}} = (\widehat{\sigma^2} + 1)^2$

Realized values (RSE = 0.3322, $\nu = 6$):

```
sigma_hat  <- 0.3322; nu<- 6
sigma2_hat <- sigma_hat^2; SSE<- nu * sigma2_hat
G_plug <- (sigma2_hat + 1)^2
G_unb  <- (SSE^2)/(nu*(nu+2)) + 2*(SSE/nu) + 1
round(c(G_plug = G_plug, G_unb = G_unb), 4)
```

So, $\widehat{G}\text{plug} \approx 1.2329$,
and $\widehat{G}\text{unb} \approx 1.2298$.

f) Model comparisons:

We compare:
(1) Full two-way with interaction, (2) Additive two-way (no interaction), (3) One-way by **type**, (4) One-way by **method**. Balanced 2×3 design, $K = 2$, so $N = 12$.

```
# Recovered ANOVA components
SS_meth <- 0.1141; SS_type <- 25.9020; SS_int <- 0.3026; SSE_full <- 0.6620
df_meth <- 1; df_type <- 2; df_int <- 2; dfE_full <- 6; N <- 12

MS_full <- SSE_full/dfE_full # Full model MSE
```

- **Test 1: Interaction (Full vs Additive)**

H_0: no interaction ($y_{ij} = 0$) vs H_1: some $y_{ij} \neq 0$

```
F_int <- (SS_int/df_int) / MS_full
p_int <- pf(F_int, df1=df_int, df2=dfE_full, lower.tail = FALSE)
```

- **Move to additive model (no interaction)**

```
SSE_add <- SSE_full + SS_int
dfE_add <- dfE_full + df_int
MS_add  <- SSE_add/dfE_add
```

- **Test 2: Method | Type in additive model**

H_0: no method effect vs H_1: method effect

```
F_meth_given_type <- (SS_meth/df_meth) / MS_add
p_meth_given_type <- pf(F_meth_given_type, df1=df_meth, df2=dfE_add,
                        lower.tail = FALSE)
```

- **Test 3: Type | Method in additive model**

H_0: no type effect vs H_1: type effect

```
F_type_given_meth <- (SS_type/df_type) / MS_add
p_type_given_meth <- pf(F_type_given_meth, df1=df_type, df2=dfE_add,
                        lower.tail = FALSE)
```

- **One-way ANOVA by type (ignoring method)**

```
SS_tot <- SS_meth + SS_type + SS_int + SSE_full
SSE_type1 <- SS_tot - SS_type
dfE_type1 <- N - 3
MS_type1  <- SSE_type1/dfE_type1
F_type_1  <- (SS_type/df_type) / MS_type1
p_type_1  <- pf(F_type_1, df1=df_type, df2=dfE_type1, lower.tail = FALSE)
```

- **One-way ANOVA by method (ignoring type)**

```
SSE_meth1 <- SS_tot - SS_meth
dfE_meth1 <- N - 2
MS_meth1  <- SSE_meth1/dfE_meth1
F_meth_1  <- (SS_meth/df_meth) / MS_meth1
p_meth_1  <- pf(F_meth_1, df1=df_meth, df2=dfE_meth1, lower.tail = FALSE)
```

```
# table
res <- data.frame(
  Test = c("Interaction (Full vs Add.)",
           "Method | Type (Add.)",
           "Type | Method (Add.)",
           "One-way Type",
           "One-way Method"),
  F_stat = c(F_int, F_meth_given_type, F_type_given_meth, F_type_1, F_meth_1),
  df1 = c(df_int, df_meth, df_type, df_type, df_meth),
  df2 = c(dfE_full, dfE_add, dfE_add, dfE_type1, dfE_meth1),
  p_value = c(p_int, p_meth_given_type, p_type_given_meth, p_type_1, p_meth_1)
)
res[ , -1] <- round(res[ , -1], 4) # Round only numeric columns
res
```

```
                       Test   F_stat df1 df2 p_value
1 Interaction (Full vs Add.)   1.3713   2   6  0.3232
2       Method | Type (Add.)   0.9463   1   8  0.3592
3       Type | Method (Add.) 107.4103   2   8  0.0000
4                One-way Type 108.0551   2   9  0.0000
5             One-way Method   0.0425   1  10  0.8409
```

Decisions ($\alpha = 0.05$):

Interaction: $F = 1.371$ ($df = 2, 6$), $p = 0.323 \Longrightarrow$ fail to reject $H_0 \longrightarrow$ no interaction.

Method | Type (additive): $F = 0.946$, $p = 0.359 \Longrightarrow$ fail to reject $H_0 \longrightarrow$ method not significant.

Type | Method (additive): $F = 107.4$, $p < 0.001 \longrightarrow$ type significant.

One-way (Type): $F = 108.1$, $p < 0.001 \longrightarrow$ type significant.

One-way (Method): $F = 0.042$, $p = 0.841 \longrightarrow$ not significant.

Conclusion: The additive two-way model is adequate (no interaction). Within it, type is highly significant and method is not. The most parsimonious adequate model is a one-way ANOVA by cheese type.

g) With known σ^2, nested linear-model tests use: $\chi^2 = \frac{SS_{\text{extra}}}{\sigma^2} \sim \chi^2_{\Delta df}$

In our balanced design, the extra sums of squares equal the ANOVA sums of squares for each effect.

```r
# Given/Recovered ANOVA components
SS_meth <- 0.1141; SS_type <- 25.9020; SS_int  <- 0.3026
df_meth <- 1; df_type <- 2; df_int  <- 2
# Known variance
sigma2_known <- 0.1
# Chi-square tests
chi2_int  <- SS_int  / sigma2_known   # Interaction
p_int     <- pchisq(chi2_int,  df=df_int,  lower.tail = FALSE)
chi2_meth <- SS_meth / sigma2_known   # Method | Type
p_meth    <- pchisq(chi2_meth, df=df_meth, lower.tail = FALSE)
chi2_type <- SS_type / sigma2_known   # Type | Method
p_type    <- pchisq(chi2_type, df=df_type, lower.tail = FALSE)
```

```r
# output
out <- data.frame(
  Effect = c("Interaction", "Method | Type", "Type | Method"),
  chi2   = c(chi2_int, chi2_meth, chi2_type),
  df     = c(df_int, df_meth, df_type),
  p      = c(p_int, p_meth, p_type)
)
out$chi2 <- round(out$chi2, 3)
out$p    <- format(out$p, digits = 3, scientific = TRUE)
out
```

```
         Effect    chi2 df        p
1   Interaction   3.026  2 2.20e-01
2 Method | Type   1.141  1 2.85e-01
3 Type | Method 259.020  2 5.68e-57
```

Decisions ($\alpha = 0.05$)
Interaction: not significant $\longrightarrow$ keep additive model.
Method | Type: not significant.
Type | Method: highly significant.

Conclusion: With known σ^2 the χ^2 tests are (weakly) more powerful than the usual F-tests, but they lead to the same model choice: moisture depends on type, not on method, with no interaction.

Section 2

a) The function to randomize the students to interfaces:

```r
# Randomized block design, skills = block levels, interfaces = treatments
# within each skill, assign each interface to exactly one student
                                                    # (random order)
```

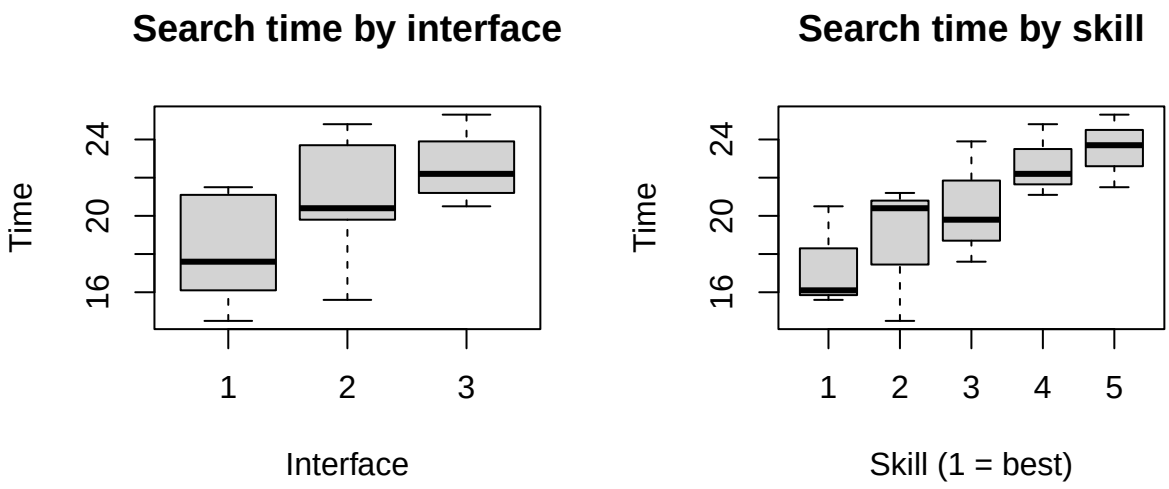
```r
randomize_interfaces <- function(skills = 1:5, interfaces = 1:3,
                                 seed = NULL) {
  if (!is.null(seed)) set.seed(seed)
  out <- data.frame(
    skill = integer(),
    student_in_skill = integer(),
    interface = integer(),
    stringsAsFactors = FALSE
  )
  for (s in skills) { # random order of interfaces for this skill
    shuffled <- sample(interfaces, length(interfaces), replace = FALSE)
    for (i in seq_along(interfaces)) {
      row <- data.frame(
        skill = s,
        student_in_skill = i,   # 1,2,3 within this skill
        interface = shuffled[i],
        stringsAsFactors = FALSE
      )
      out <- rbind(out, row)
    }
  }
  out
}
```

b) Graphical summaries of the data:

```r
par(mfrow=c(1,2))  # 1 row, 2 columns

# Boxplot 1
boxplot(time ~ interface, data = search,
        main = "Search time by interface",
        xlab = "Interface",
        ylab = "Time")

# Boxplot 2
boxplot(time ~ skill, data = search,
        main = "Search time by skill",
        xlab = "Skill (1 = best)",
        ylab = "Time")
```

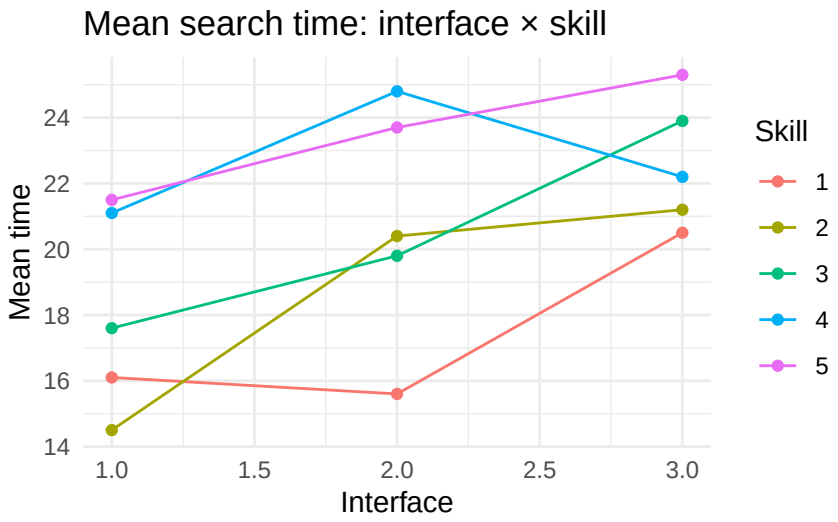


Comment - Search time by interface: The plot suggests that interface 1 has the shortest search times on average, while interface 3 is the slowest. This indicates a possible effect of the interface on performance.

Comment - Search time by skill: Students with better computer handling (skill = 1) are clearly faster. As the skill level number increases, the average search time also increases, showing a strong effect of skill.

```
# load ggplot2
library(ggplot2)

# mean plot: interface × skill
ggplot(search, aes(x = interface, y = time,
                   color = factor(skill), group = skill)) +
  stat_summary(fun = mean, geom = "line") +
  stat_summary(fun = mean, geom = "point") +
  labs(title = "Mean search time: interface × skill",
       x = "Interface",
       y = "Mean time",
       color = "Skill") +
  theme_minimal()
```



Comment - Interface x skill: The lines are not completely parallel, suggesting some interaction between interface and skill. In other words, the difference between interfaces may depend on the student's skill level.

c) ANOVA test for interface effect:
Response: `time` (continuous)
Treatment factor: `interface` (3 levels)
Block factor: `skill` (5 levels)
We want a two-way ANOVA without interaction (since in randomized block design, interaction isn't usually the main focus)

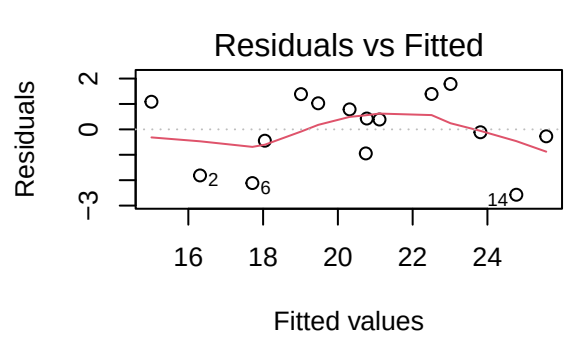
Hypotheses: $H_0 : \mu_{\text{interface 1}} = \mu_{\text{interface 2}} = \mu_{\text{interface 3}}$ H_1 : At least one mean differs

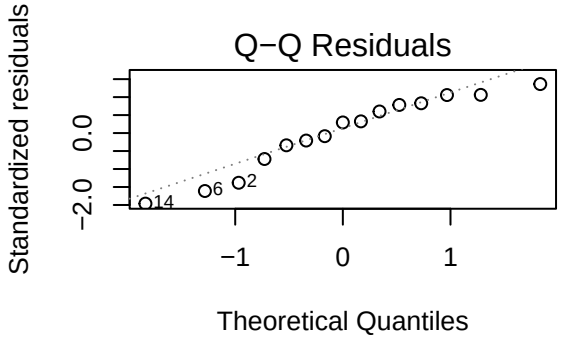
```
search$interface <- factor(search$interface) # Convert to factors
search$skill <- factor(search$skill)
fit <- aov(time ~ interface + skill, data = search) # ANOVA model
summary(fit) # ANOVA table
```

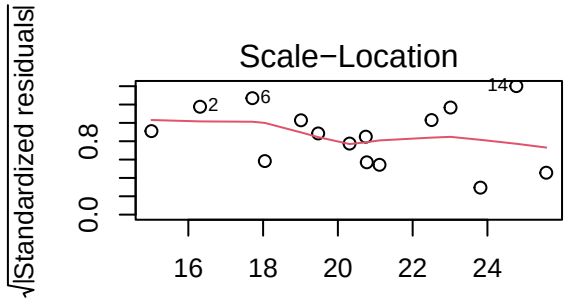
```
            Df Sum Sq Mean Sq F value Pr(>F)
interface    2  50.47  25.233   7.824 0.0131 *
skill        4  80.05  20.013   6.205 0.0142 *
Residuals    8  25.80   3.225
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

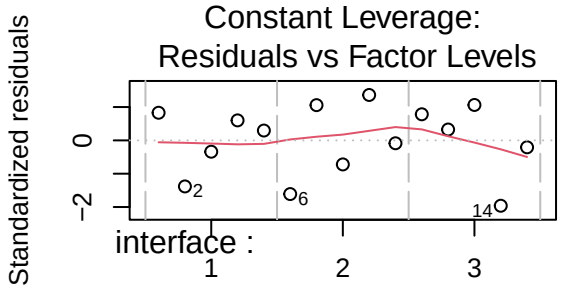
Result:
The ANOVA shows a significant effect of `interface` ($F = 7.82, p = 0.013$) and also of `skill` ($F = 6.21, p = 0.014$).
Therefore, we reject the null hypothesis that all interfaces have the same mean search time. Both interface and skill influence search time.

```r
par(mfrow = c(2,2))
plot(fit)                              # 1) Residuals vs Fitted  2) QQ  3) Scale-Location
```










```r
shapiro.test(residuals(fit))          # Normality of errors (p > .05 -> OK)
```

```
    Shapiro-Wilk normality test

data:  residuals(fit)
W = 0.93092, p-value = 0.2817
```

```r
par(mfrow = c(1,1))
```

Assumption checks
The residual diagnostics do not indicate violations of the ANOVA model assumptions. The Residuals vs Fitted plot shows no systematic pattern, suggesting that the variance is approximately constant across fitted values. The Scale–Location plot also does not show evidence of heteroscedasticity. The Q–Q plot is close to the reference line, and the Shapiro–Wilk test does not reject normality of the residuals ($p = 0.28 > 0.05$). Therefore, the normality and homoscedasticity assumptions appear to be reasonably satisfied, and the F–test results are valid.

(d) Interface preference and predictions

From the ANOVA in part (c), both `interface` and `skill` are significant.
We now check which interface is preferable, how skill affects search time, and make predictions for specific cases.

```
tapply(search$time, search$interface, mean) # Estimated means per interface
tapply(search$time, search$skill, mean) # Estimated means per skill
fit <- aov(time ~ interface + skill, data = search) # Prediction model (c)
```

```
# Prediction for skill 4, interface 3
newdata1<-data.frame(interface = factor(3, levels=levels(search$interface)),
                     skill = factor(4, levels=levels(search$skill)))
predict(fit, newdata1)
# Prediction for skill 1, interface 2
newdata2<-data.frame(interface = factor(2, levels=levels(search$interface)),
                     skill = factor(1, levels=levels(search$skill)))
predict(fit, newdata2)
```

Result:
- Mean times per interface show that `Interface 1` is fastest (≈ 18.2), followed by `Interface 2` (≈ 20.9), and `Interface 3` is slowest (≈ 22.6).
$\Longrightarrow$ Therefore, `Interface 1` is preferable.
- Mean times per skill show a clear increase from `skill 1` (≈ 17.4) to `skill 5` (≈ 23.5).
$\Longrightarrow$ Search time increases as skill decreases (higher skill number = slower).
- Predicted times:
`Skill 4, interface 3`: ≈ 24.7 seconds \$ `Skill 1, interface 2`: ≈ 17.7 seconds.

e) One-way ANOVA ignoring skill

Hypotheses: $H_0 : \mu_{\text{interface 1}} = \mu_{\text{interface 2}} = \mu_{\text{interface 3}}$ H_1 : At least one mean differs

```
fit_oneway <- aov(time ~ interface, data = search)
summary(fit_oneway) # One-way ANOVA ignoring skill
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
interface    2  50.47  25.233   2.861 0.0964 .
Residuals   12 105.85   8.821
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Result:
The one-way ANOVA (ignoring skill) gives:

- `interface`: $F \approx 2.86$, $p \approx 0.096$ (not significant at 5%).

Interpretation:
- If we ignore skill, the test suggests no significant difference between interfaces.
- However, this is misleading, because in part (c) we saw that interface was significant after controlling for skill.

- Skill is an important block factor (strong effect on time). Ignoring it increases the residual variation, which makes the interface effect harder to detect.

Conclusion:
Performing the interface-only ANOVA (ignoring skill) is not meaningful for this dataset. The correct approach is the randomized block design (including skill), as done in part (c).

Section 3

Add to the data frame the variable weightlost:

```
diet_df$weightlost<-diet_df$preweight - diet_df$weight6weeks
```

Disregarding the columns: person, preweight and weight6weeks:

```
diet_df<-subset(diet_df, select = -c(person, preweight, weight6weeks))
```

a) One-way ANOVA: effect of diet on weight loss

```
diet_df$diet<-as.factor(diet_df$diet)
diet_one_way<-aov(weightlost~diet, data = diet_df)
bartlett.test(weightlost ~ diet, data = diet_df) # homogeneity of variances
```

```
    Bartlett test of homogeneity of variances

data:  weightlost by diet
Bartlett's K-squared = 0.33745, df = 2, p-value = 0.8447
```

Bartlett's test: $p = 0.84 \Longrightarrow$ homogeneity of variances holds.

```
summary(diet_one_way)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
diet         2   71.1   35.55   6.197 0.00323 **
Residuals   75  430.2    5.74
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results $F(2, 75) = 6.20$, $p = 0.003 < 0.01$
$\Longrightarrow$ The type of diet has a statistically significant effect on mean weight loss.

Group means

```
tapply(diet_df$weightlost, diet_df$diet, mean)
```

Diet 1: 3.30 kg Diet 2: 3.03 kg Diet 3: 5.15 kg

All diets led to average weight loss; Diet 3 showed the highest.

Post-hoc (Tukey)

```r
TukeyHSD(diet_one_way)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weightlost ~ diet, data = diet_df)

$diet
          diff        lwr      upr     p adj
2-1 -0.2740741 -1.8806155 1.332467 0.9124737
3-1  1.8481481  0.2416067 3.454690 0.0201413
3-2  2.1222222  0.5636481 3.680796 0.0047819
```

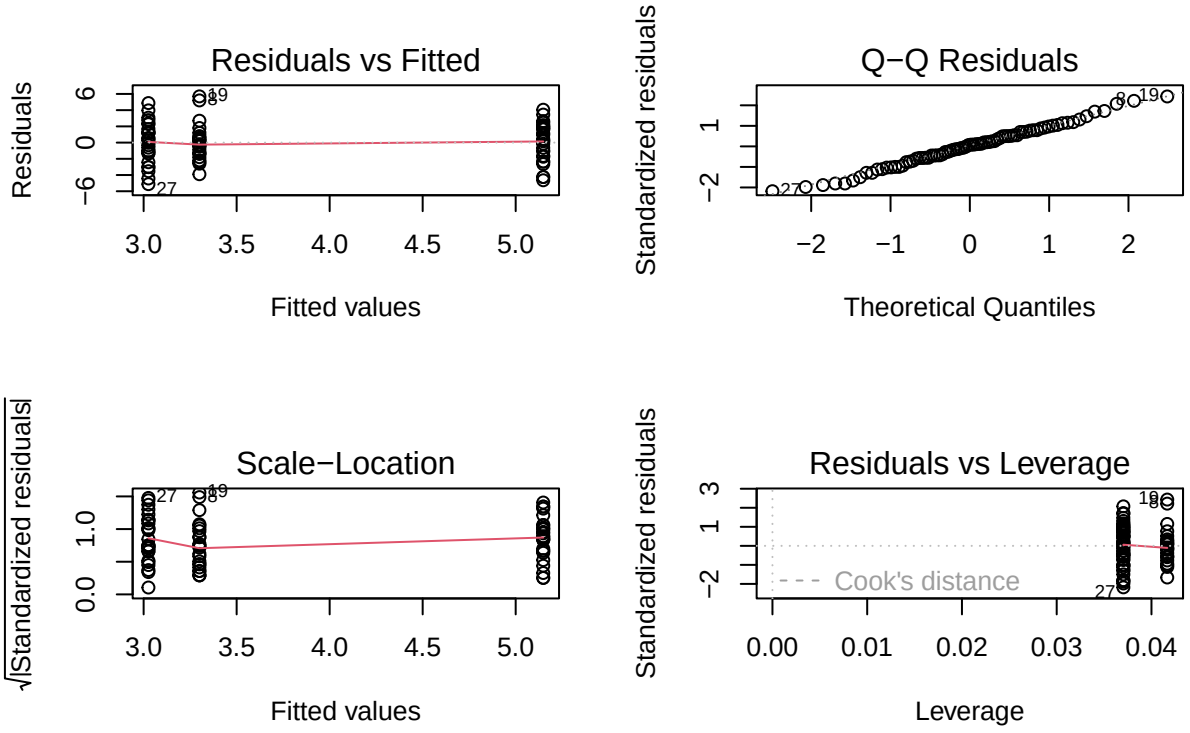
Diet 3 > Diet 1 ($p = 0.020$) Diet 3 > Diet 2 ($p = 0.0048$) Diet 1 ≈ Diet 2 ($p = 0.91$)

Conclusion: Diet 3 is significantly more effective than Diets 1 and 2.

Relation to t-test

With two groups, ANOVA reduces to a two-sample t-test ($F = t^2$).
Here, with three groups, one-way ANOVA was required, followed by pairwise tests.

```r
# Residual diagnostics plots
par(mfrow = c(2,2))
plot(diet_one_way)
```



```r
par(mfrow = c(1,1))
# Normality of residuals
shapiro.test(residuals(diet_one_way))
```

```
    Shapiro-Wilk normality test

data:  residuals(diet_one_way)
W = 0.99208, p-value = 0.9161
```

Assumption check
The Residuals vs Fitted and Scale–Location plots do not show any systematic pattern or funnel shape, suggesting that the variance is approximately constant across the diet groups. The Q–Q plot is close to the reference line, indicating no substantial deviation from normality. The Shapiro–Wilk test supports this conclusion ($p = 0.9161 > 0.05$), so we do not reject normality. No influential observations are suggested by the Residuals vs Leverage plot. Therefore, the assumptions of one-way ANOVA (normality and homoscedasticity of errors) appear to be reasonably satisfied, and the F-test result is valid.

b) Two-way ANOVA: diet, gender, and interaction

```r
diet_df$diet<-as.factor(diet_df$diet)
diet_df$gender<-as.factor(diet_df$gender)
diet_two_way<-aov(weightlost~gender+diet+gender*diet, data = diet_df)
summary(diet_two_way)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
gender       1    0.3   0.278   0.052 0.82062
diet         2   60.4  30.209   5.619 0.00546 **
gender:diet  2   33.9  16.952   3.153 0.04884 *
Residuals   70  376.3   5.376
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
2 observations deleted due to missingness
```

Results Gender: $F(1, 70) = 0.05$, $p = 0.82 \Longrightarrow$ not significant
Diet: $F(2, 70) = 5.62$, $p = 0.0055 \Longrightarrow$ significant
Diet $\times$ Gender interaction: $F(2, 70) = 3.15$, $p = 0.049 \Longrightarrow$ indicating a significant interaction.

Interpretation: The significant diet $*$ gender interaction indicates that the impact of diet differs between men and women.

c) ANCOVA: adding height and age
From (a): diet had a clear effect on weight loss.
From (b): the diet $*$ gender interaction was significant $\Longrightarrow$ evidence that men and women respond differently to diets.
When adding height and age as covariates, our goal is to control for their possible influence

while retaining the key factors identified earlier (diet, gender, and their interaction).

We extend the model to include continuous covariates:

$$\text{weight loss} \sim \text{diet} + \text{gender} + \text{height} + \text{age} + \text{diet} \times \text{gender}$$

```
# --- Center once (plain subtraction avoids type issues)
age_mean    <- mean(diet_df$age,    na.rm = TRUE)
height_mean <- mean(diet_df$height, na.rm = TRUE)
diet_df$age_c    <- diet_df$age    - age_mean
diet_df$height_c <- diet_df$height - height_mean
diet_df$diet   <- factor(diet_df$diet)
diet_df$gender <- factor(diet_df$gender)

# --- Candidate set: at most ONE interaction
forms <- list(
  m0  = weightlost ~ diet + gender + height_c + age_c,
  mDG = weightlost ~ diet*gender + height_c + age_c,
  mDA = weightlost ~ diet*age_c   + gender + height_c,
  mDH = weightlost ~ diet*height_c+ gender + age_c,
  mGA = weightlost ~ diet + gender*age_c   + height_c,
  mGH = weightlost ~ diet + gender*height_c+ age_c
)

fits <- lapply(forms, \(f) lm(f, data = diet_df))
aic_tbl <- data.frame(
  model = names(fits),
  k     = sapply(fits, \(m) length(coef(m))),
  AIC   = sapply(fits, AIC),
  row.names = NULL
)
aic_tbl <- aic_tbl[order(aic_tbl$AIC), ]

# Nested F only where valid (interaction adds terms to m0)
cmp_DG <- anova(fits$m0, fits$mDG)

# Choose final model = AIC winner
best_name <- aic_tbl$model[1]
m_final   <- fits[[best_name]]
best_name; aic_tbl; cmp_DG
```

```
[1] "mDG"
```

```
  model k      AIC
2   mDG 8 354.7980
5   mGA 7 357.1822
```

```
1    m0 6 357.5882
4   mDH 8 358.4031
6   mGH 7 359.0815
3   mDA 8 361.0773
```

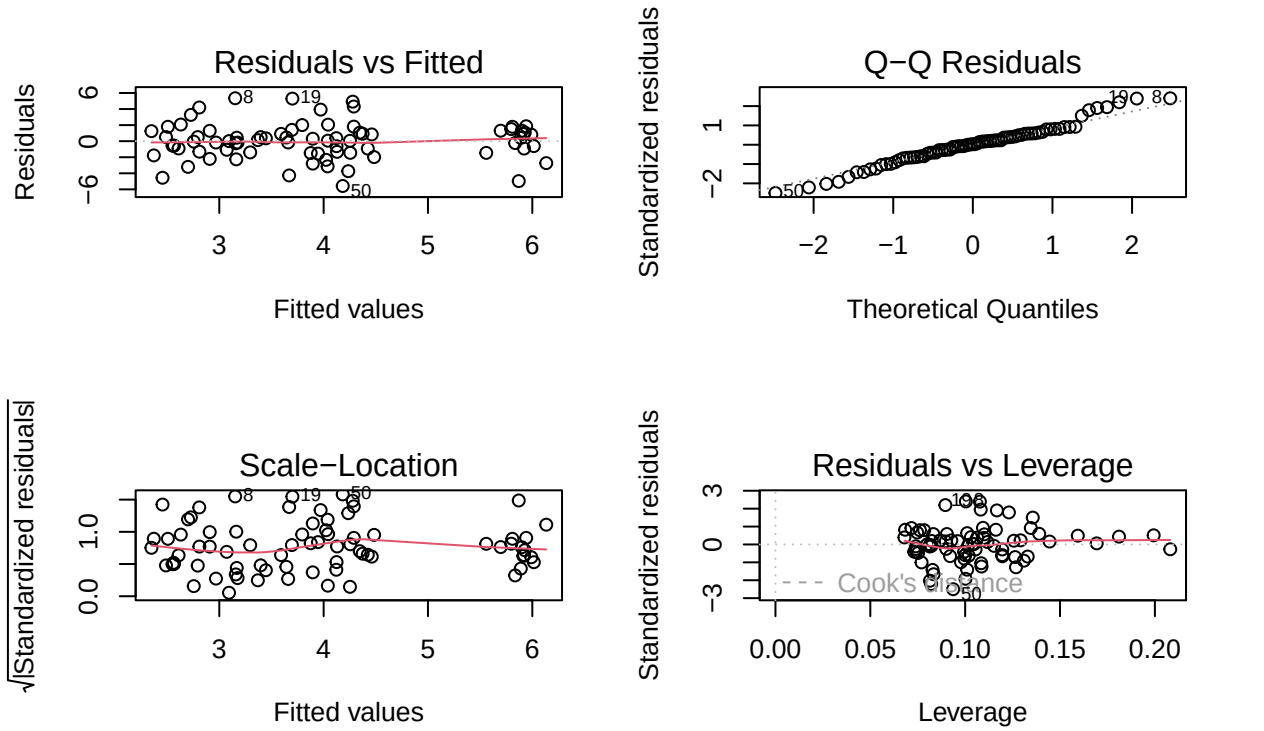
```
Analysis of Variance Table

Model 1: weightlost ~ diet + gender + height_c + age_c
Model 2: weightlost ~ diet * gender + height_c + age_c
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     70 409.02
2     68 374.06  2    34.959 3.1776 0.04794 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Winner model (interaction + centered covariates)
m_final <- lm(weightlost ~ diet*gender + height_c + age_c, data = diet_df)

# --- Diagnostics ---

# Residual plots
par(mfrow = c(2,2))
plot(m_final)
```



```r
par(mfrow = c(1,1))

# Normality (Shapiro-Wilk test)
shapiro.test(residuals(m_final))
```

```
    Shapiro-Wilk normality test

data:  residuals(m_final)
W = 0.97761, p-value = 0.1986
```

Assumptions
The residual vs fitted and scale–location plots do not show systematic structure, the Q–Q plot is approximately linear, and the Shapiro–Wilk test does not reject normality ($p > 0.05$). Therefore, the model assumptions appear reasonably satisfied.

Interpretation
We compare the additive model with a model including the diet×gender interaction. The nested F-test (ANOVA) shows that adding the interaction significantly improves the model fit ($p = 0.048$). Moreover, AIC is lowest for the interaction model.

Diet has a strong effect on weight loss, while gender alone does not. However, the interaction indicates that the effect of diet differs between men and women. Height and age are not individually significant, but are kept as adjustment covariates. Therefore, our final model is:

$$\text{weightlost} \sim \text{diet} \times \text{gender} + \text{height}_c + \text{age}_c.$$

d) Predictions for an "average person"

```r
age_av    <- mean(diet_df$age, na.rm = TRUE)
height_av <- mean(diet_df$height, na.rm = TRUE)

grid <- expand.grid(
  diet   = levels(diet_df$diet),
  gender = levels(diet_df$gender)
)

grid$age_c    <- age_av - mean(diet_df$age)
grid$height_c <- height_av - mean(diet_df$height)

grid$pred <- round(predict(m_final, newdata = grid), 1)
grid
```

```
  diet gender age_c height_c pred
1    1      0     0        0  3.1
2    2      0     0        0  2.6
```

```
3    3      0     0        0  5.8
4    1      1     0        0  3.6
5    2      1     0        0  4.3
6    3      1     0        0  4.2
```

Interpretation:
For an average person (height ≈ 170.8 cm, age ≈ 39.2 years), all diets lead to weight loss. The largest expected weight loss occurs for females following Diet 3 (≈ 5.8 kg). Therefore, Diet 3 is the most effective, particularly for women.

Section 4

In this exercise we apply the LASSO method to the Birthweights dataset, with the goal of identifying relevant predictors of newborn birthweight. We compare the LASSO model against the full linear regression model. The comparison is based on prediction error (MSE) and interpretability in terms of selected predictors.

Lasso for Birthweights

```r
# dataset: birthw_df ; response: Birthweight
# x = all predictors, y = response
x <- as.matrix(birthw_df[ , !(names(birthw_df) %in% c("Birthweight", "ID"))])
y <- birthw_df$Birthweight
```

The variable ID is an identifier and contains no predictive information about birthweight. Therefore, it must be removed before fitting the models.

```r
set.seed(42)
train <- sample(1:nrow(x), 0.67 * nrow(x))  # use ~2/3 for training
x.train <- x[train, ];  y.train <- y[train]  # training data
x.test  <- x[-train,];  y.test  <- y[-train] # test data
```

Prediction by the linear model (trained on train rows)

```r
lm.model <- lm(Birthweight ~ ., data = birthw_df, subset = train,
               na.action = na.exclude)
y.predict.lm <- predict(lm.model, newdata = birthw_df[-train, ])
mse.lm <- mean((y.test - y.predict.lm)^2)
```

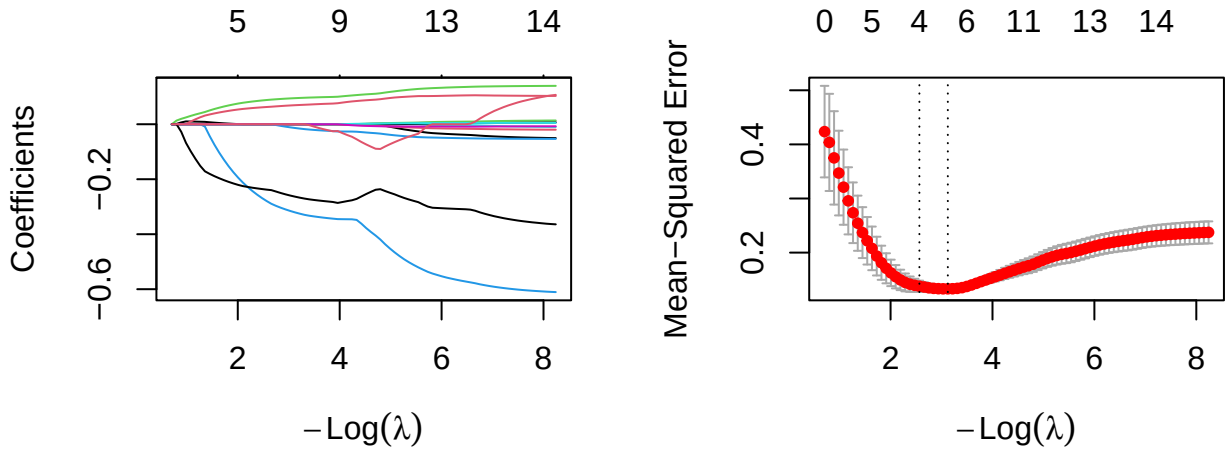
The full linear regression model uses all available predictors and yields
a test-set MSE of 0.398.
This serves as our baseline for evaluating whether LASSO can achieve better generalization performance with fewer variables.

Lasso application for variable selection and prediction

```r
library(glmnet)

set.seed(42)
```

```r
# Plain lasso fit on train (alpha=1)
lasso.model <- glmnet(x.train, y.train, alpha = 1)
# Cross-validated lasso (type.measure = "mse", 5-fold to match example)
lasso.cv <- cv.glmnet(x.train, y.train, alpha = 1, type.measure = "mse",
                      nfolds = 5)
```




```r
# lambdas
lambda.min  <- lasso.cv$lambda.min          # min CV error
lambda.1se  <- lasso.cv$lambda.1se          # 1-SE (more regularized)
```

The cross-validation yielded $\lambda_{min} = 0.044$, which minimizes the cross-validated error and $\lambda_{1se} = 0.077$, which gives a more regularized model within one standard error of the minimum. The latter is typically preferred because it offers a simpler model with comparable prediction accuracy.

Note: In practice, the λ_{1se} model is chosen for its simplicity, as it achieves nearly the same predictive performance as λ_{min} while using fewer predictors.

```r
# Coefficients at the two lambdas
coef.min <- as.matrix(coef(lasso.model, s = lambda.min))
coef.1se <- as.matrix(coef(lasso.model, s = lambda.1se))
sel.min <- setdiff(rownames(coef.min)[coef.min[, 1] != 0], "(Intercept)")
sel.1se <- setdiff(rownames(coef.1se)[coef.1se[, 1] != 0], "(Intercept)")
```

`Selected variables at each lambda`

Lambda	Selected predictors
λ_{min}	Headcirc, Gestation, Smoker, Fedyrs, Lowbwt
λ_{1se}	Headcirc, Gestation, Smoker, Lowbwt

Both LASSO models exclude most predictors and retain only a few key variables. Head circumference and gestation length consistently appear, which is medically plausible since they are directly linked to fetal growth. Smoking status and low birthweight indicators also remain, underlining their strong influence. At λ_{min}, the father's years in education (Fedyrs) enters, but disappears again at λ_{1se}, suggesting it has only marginal predictive power.

```
# Test-set predictions and MSEs for both lambdas
lasso.pred.min <- as.numeric(predict(lasso.model, s = lambda.min,
                                     newx = x.test))
lasso.pred.1se <- as.numeric(predict(lasso.model, s = lambda.1se,
                                     newx = x.test))
mse.min  <- mean((y.test - lasso.pred.min)^2)
mse.1se  <- mean((y.test - lasso.pred.1se)^2)
```

The test-set MSE is 0.162 at $\lambda_{\min}$ and 0.154 at λ_{1se}
Both LASSO models clearly outperform the full linear regression in terms of prediction error (≈ 0.16 and ≈ 0.15 vs. 0.398). More importantly, the LASSO approach achieves this accuracy with only four to five predictors, whereas the full linear model includes all the available variables."

Summary: The LASSO models achieve test-set MSEs of 0.162 ($\lambda_{\min}$) and 0.154 (λ_{1se}), both substantially lower than that of the full linear model(0.398). At the same time, LASSO provides a more parsimonious specification: only a small set of predictors (Headcirc, Gestation, Smoker, and Lowbwt, with Fedyrs included only at $\lambda_{\min}$) are retained. This shows that LASSO not only maintains or improves predictive accuracy but also improves interpretability by excluding irrelevant variables.

Conclusion

In summary, this project applied a range of linear modeling techniques—including one-way and two-way ANOVA, randomized block design, ANCOVA, and LASSO regression—to examine how different factors influence the response variables in each dataset. The analyses were performed in R and supported by diagnostic checks to assess model assumptions and guide model selection.

Across the exercises, the key findings highlighted the importance of choosing an appropriate model structure, evaluating interaction effects when relevant, and comparing models based on both statistical fit and interpretability. The use of LASSO demonstrated how regularization can improve predictive performance when many predictors are present. Overall, the work illustrates how statistical modeling can be used systematically to extract meaningful conclusions and support data-driven decision making.