Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
ΔΠΜΣ Επιστήμη Δεδομένων & Μηχανική Μάθηση

Υπολογιστική Στατιστική & Στοχαστική Βελτιστοποίηση

Εξαμηνιαία Εργασία Μαθήματος

Συγγραφή:
Λυδία Ιωάννα Κολίτση
ΑΜ: 03400252
lydiannakolitsi@mail.ntua.gr

Διδάσκοντες:
Δημήτριος Φουσκάκης
Μιχαήλ Λουλάκης
Κωνσταντίνα Χαρμπή

19 Ιουνίου 2025

Περιεχόμενα

Κατάλογος σχημάτων

Άσκηση 1

Ερώτημα 1 (α) i: Μη Παραμετρική Παλινδρόμηση Nadaraya-Watson με LOOCV

Δίνεται το σύνολο δεδομένων `data1pairs.rds`, το οποίο αποτελείται από 200 ζεύγη παρατηρήσεων (x_i, y_i) $(i = 1, ..., 200)$ τυχαίου δείγματος $(X_1, Y_1), ..., (X_{200}, Y_{200})$ και ζητείται η εφαρμογή μη παραμετρικής παλινδρόμησης. Ειδικότερα, χρησιμοποιείται η συνάρτηση `ksmooth` της `R` για τον υπολογισμό της εκτίμησης που δίνει η μέθοδος Nadaraya-Watson με Gaussian πυρήνα (kernel) και χρήση leave-one-out cross-validation (LOOCV) με κριτήριο το μέσο τετραγωνικό σφάλμα για τον υπολογισμό του βέλτιστου πλάτους h_x.

Στόχος είναι η εκτίμηση της συνάρτησης παλινδρόμησης $m(x) = E[Y|X = x]$, δηλαδή της αναμενόμενης τιμής της μεταβλητής Y για κάθε τιμή της X, χωρίς να προϋποτίθεται κάποια συγκεκριμένη μαθηματική μορφή για τη μεταξύ τους σχέση. Η εκτίμηση επιτυγχάνεται με τη μέθοδο Nadaraya-Watson, η οποία βασίζεται σε τοπικούς σταθμισμένους μέσους όρους. Η βασική ιδέα είναι ότι οι παρατηρήσεις που βρίσκονται «κοντά» στην τιμή x συμβάλλουν περισσότερο στην εκτίμηση της $m(x)$, ενώ οι πιο απομακρυσμένες έχουν μικρότερη βαρύτητα. Αυτή η λογική υλοποιείται με χρήση συναρτήσεων πυρήνα (kernel functions), οι οποίες αποδίδουν βάρη στις παρατηρήσεις ανάλογα με την απόστασή τους από το x. Ο εκτιμητής Nadaraya-Watson δίνεται από τη σχέση [1]:

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right) y_i}{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right)} = \sum_{i=1}^{n} w_i y_i.$$

Η συνάρτηση $K_x(\cdot)$ αποτελεί τον πυρήνα, ο οποίος στη συγκεκριμένη περίπτωση είναι ο Gaussian:

$$K_x(u) \;=\; \frac{1}{\sqrt{2\pi}}\,\exp\!\left(-\frac{u^2}{2}\right).$$

Η παράμετρος h_x (bandwidth) ρυθμίζει τον βαθμό εξομάλυνσης: μικρές τιμές h_x οδηγούν σε εκτιμήσεις που ακολουθούν πιο πιστά τα δεδομένα, ενώ μεγαλύτερες τιμές οδηγούν σε πιο ομαλές εκτιμήσεις. Η επιλογή της κατάλληλης τιμής h_x γίνεται με τη μέθοδο Leave-One-Out Cross-Validation (LOOCV) [2], όπου για κάθε υποψήφια τιμή του h_x, γίνεται επαναληπτικά αφαίρεση μιας παρατήρησης (x_i, y_i) και προσαρμογή του μοντέλου Nadaraya-Watson στις υπόλοιπες $n - 1$ παρατηρήσεις,

$$\widehat{m}_{(-i)}(x) = \frac{\displaystyle\sum_{\substack{j=1\\ j\neq i}}^{n} K\Big(\frac{x-x_j}{h_x}\Big)\, y_j}{\displaystyle\sum_{\substack{j=1\\ j\neq i}}^{n} K\Big(\frac{x-x_j}{h_x}\Big)}.$$

Γίνεται πρόβλεψη της τιμής της εξαρτημένης μεταβλητής στο σημείο x_i που αφαιρέθηκε, δηλαδή υπολογίζεται το $\hat{y}_{-i} = \hat{m}_{-i}x_i$. Το σφάλμα πρόβλεψης για την παρατήρηση i είναι $y_i - \hat{y}_{-i}$. Το συνολικό κριτήριο απόδοσης για το συγκεκριμένο h_x είναι το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error):

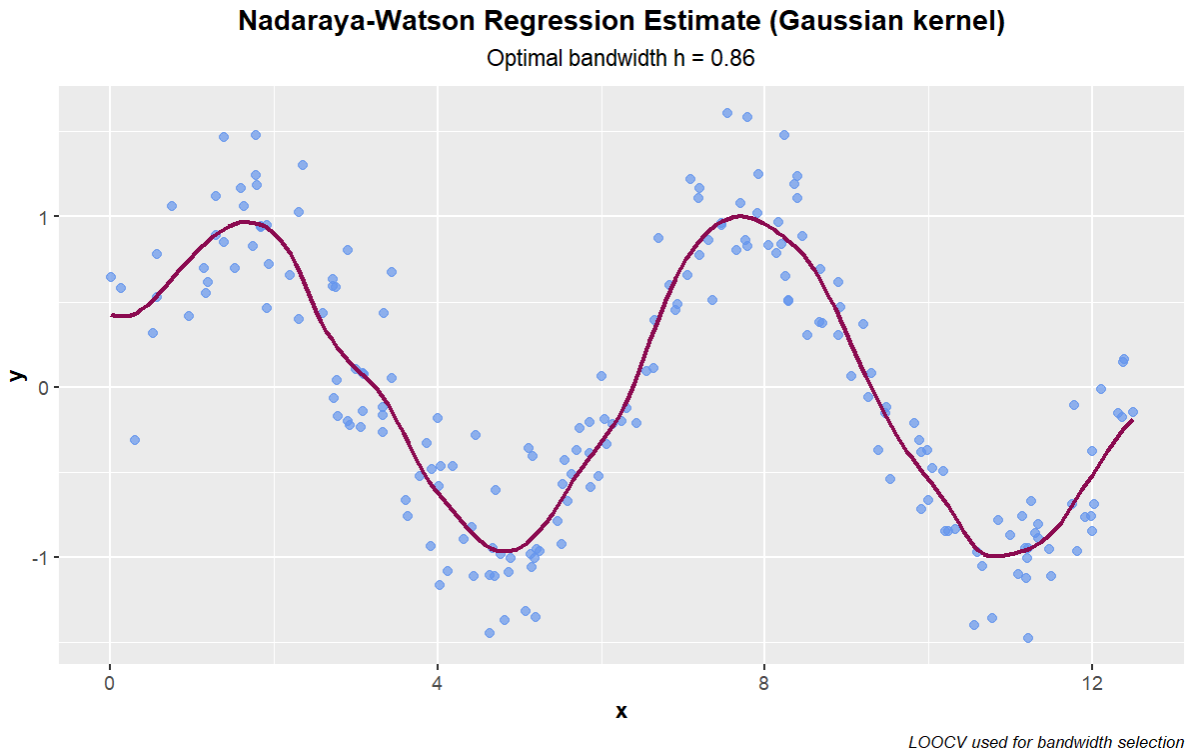
$$\mathrm{MSE}_{\mathrm{LOOCV}}(h_x) \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \widehat{y}_{i,-i}\big)^2.$$

Η διαδικασία αυτή επαναλαμβάνεται για ένα εύρος υποψήφιων τιμών του h_x και τελικά επιλέγεται η τιμή που ελαχιστοποιεί το $\mathrm{MSE}_{\mathrm{LOOCV}}(h_x)$.

Ακολούθως παρουσιάζεται ο κώδικας που υλοποιεί την εκτίμηση Nadaraya–Watson με κανονικό πυρήνα. Ορίζεται πλέγμα τιμών για το bandwidth h_x μέσω της μεταβλητής `h_grid` και εφαρμόζεται Leave-One-Out Cross-Validation για κάθε τιμή: σε κάθε βήμα, εξαιρείται το σημείο i και εκτιμάται η τιμή του y_i από τα υπόλοιπα

σημεία με χρήση της συνάρτησης `ksmooth`. Για κάθε h_x, υπολογίζεται το μέσο τετραγωνικό σφάλμα (MSE) και αποθηκεύεται στον πίνακα `cv_mse`. Η βέλτιστη τιμή `best_h` επιλέγεται ως εκείνη που ελαχιστοποιεί το `cv_mse`. Η τελική καμπύλη παλινδρόμησης εκτιμάται ξανά μ την `ksmooth` στο πλήρες σύνολο δεδομένων και τα αποτελέσματα απεικονίζονται με χρήση `ggplot2`.

```r
# Load required libraries
library(ggplot2)
library(dplyr)

# 1. Read the data
data <- readRDS("data1pairs.rds")
x <- data[, 1]
y <- data[, 2]
n <- length(x)

# 2. Define a grid of bandwidth values
h_grid <- seq(0.05, 1, by = 0.01)
cv_mse <- numeric(length(h_grid))

# 3. Perform Leave-One-Out Cross-Validation for each h
for (j in seq_along(h_grid)) {
  h <- h_grid[j]
  predictions <- numeric(n)

  for (i in 1:n) {
    # Leave one observation out
    x_train <- x[-i]
    y_train <- y[-i]

    # Estimate y_i using ksmooth on the rest of the data
    pred <- ksmooth(x_train, y_train, kernel = "normal", bandwidth = h, x.points =
    x[i])
    predictions[i] <- pred$y
  }

  # Calculate MSE for current h
  cv_mse[j] <- mean((y - predictions)^2)
}

# 4. Select the optimal bandwidth (h that minimizes CV MSE)
best_h <- h_grid[which.min(cv_mse)]
cat("Optimal bandwidth h:", best_h, "\n")

# 5. Use ksmooth with the optimal h to compute final fitted values
fit <- ksmooth(x, y, kernel = "normal", bandwidth = best_h)

# 6. Prepare data for plotting
df_points <- data.frame(x = x, y = y)
df_fit <- data.frame(x = fit$x, y_hat = fit$y)


```



Σχήμα 1. *Εκτίμηση συνάρτησης παλινδρόμησης με τη μέθοδο Nadaraya-Watson και βέλτιστο bandwidth* $h = 0.86$ *(Gaussian πυρήνας, επιλογή μέσω Leave-One-Out Cross-Validation).*

```
# 7. Plot using ggplot2
ggplot() +
  geom_point(data = df_points, aes(x = x, y = y), color = "cornflowerblue", size =
    2.5, alpha = 0.7) +
  geom_line(data = df_fit, aes(x = x, y = y_hat), color = "deeppink4", linewidth =
    1.2) +
  labs(
    title = "Nadaraya-Watson Regression Estimate (Gaussian kernel)",
    subtitle = paste("Optimal bandwidth h =", round(best_h, 3)),
    x = "x",
    y = "y",
    caption = "LOOCV used for bandwidth selection"
  ) +
  theme_gray(base_size = 14)   +
  theme(
    #panel.border = element_rect(color = "black", fill = NA, linewidth = 0.5),
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5, margin = margin(b = 10)),
    axis.title = element_text(face = "bold"),
    plot.caption = element_text(size = 10, face = "italic", hjust = 1))
```

Στο Σχήμα 1 παρουσιάζονται οι παρατηρήσεις (x_i, y_i), καθώς και η εκτιμηθείσα καμπύλη παλινδρόμησης μέσω της μεθόδου Nadaraya-Watson με Gaussian πυρήνα. Η βέλτιστη τιμή πλάτους υπολογίστηκε με Leave-One-Out Cross-Validation ως $h_x = 0.86$. Φαίνεται ότι η καμπύλη Nadaraya-Watson ακολουθεί ικανοποιητικά τη γενική δομή των δεδομένων, αποτυπώνοντας τη μη γραμμική συσχέτιση ανάμεσα στις μεταβλητές, χωρίς να υπερπροσαρμόζεται στον τυχαίο θόρυβο των σημείων.

Ερώτημα 1 (α) ii: Αποδοτικότερη Λύση: Ο Τύπος του PRESS

Στο προηγούμενο ερώτημα, το βέλτιστο πλάτος h_x για τον εκτιμητή Nadaraya–Watson προσδιορίστηκε μέσω της μεθόδου Leave-One-Out Cross-Validation, η οποία απαιτεί την επαναπροσαρμογή του μοντέλου n φορές (όπου n το πλήθος των παρατηρήσεων) για κάθε υποψήφια τιμή του h_x. Αν και ακριβής, αυτή η προσέγγιση είναι υπολογιστικά δαπανηρή για μεγάλα σύνολα δεδομένων.

Ευτυχώς, για γραμμικούς ομαλοποιητές (linear smoothers) όπως ο εκτιμητής Nadaraya-Watson, υπάρχει ένας αποδοτικός αναλυτικός τύπος για τον υπολογισμό του σφάλματος LOOCV, ο οποίος απαιτεί μία μόνο προσαρμογή του μοντέλου σε ολόκληρο το σύνολο δεδομένων [2, 3].

Ένας εκτιμητής θεωρείται γραμμικός ομαλοποιητής εάν το διάνυσμα των προσαρμοσμένων τιμών, $\hat{y}$, μπορεί να εκφραστεί ως γραμμικός μετασχηματισμός του διανύσματος των παρατηρούμενων τιμών y:

$$\hat{y} = Sy,$$

όπου ο S είναι ένας πίνακας διαστάσεων $n \times n$ που ονομάζεται πίνακας ομαλοποίησης (smoother matrix), ο οποίος είναι το ανάλογο του πίνακα προβολής (hat matrix) της παραμετρικής γραμμικής παλινδρόμησης. Τα στοιχεία του πίνακα S εξαρτώνται αποκλειστικά από τις τιμές των ανεξάρτητων μεταβλητών x_i και την παράμετρο ομαλοποίησης h_x. Συγκεκριμένα, για τον εκτιμητή Nadaraya-Watson, το στοιχείο S_{ij} του πίνακα αντιστοιχεί στο βάρος $w_j(x_i)$ που αποδίδεται στην παρατήρηση y_j κατά τον υπολογισμό της προσαρμοσμένης τιμής στο σημείο x_i:

$$S_{ij} = w_j(x_i) = \frac{K(\frac{x_i - x_j}{h_x})}{\sum_{k=1}^{n} K(\frac{x_i - x_k}{h_x})}.$$

Ουσιαστικά, το υπόλοιπο της διασταυρούμενης επικύρωσης $(y_i - \hat{y}_{-i})$ με το υπόλοιπο της κανονικής προσαρμογής $(y_i - \hat{y}_i)$ συνδέεται μέσω της σχέσης:

$$y_i - \hat{y}_{-i} = \frac{y_i - \hat{y}_i}{1 - S_{ii}}$$

Το διαγώνιο στοιχείο S_{ii} του πίνακα ομαλοποίησης ερμηνεύεται ως ο βαθμός "αυτο-επιρροής" (leverage) της παρατήρησης y_i στον υπολογισμό της δικής της προσαρμοσμένης τιμής $\hat{y}_i$. Ουσιαστικά, ο παραπάνω τύπος διορθώνει το υπόλοιπο της κανονικής προσαρμογής, διαιρώντας το με έναν παράγοντα που αντισταθμίζει τη μεροληψία που προκύπτει από τη χρήση της ίδιας της παρατήρησης y_i στην εκτίμηση $\hat{y}_i$.

Συνεπώς, το κριτήριο MSE μπορεί να εκφραστεί χρησιμοποιώντας μόνο ποσότητες που υπολογίζονται από μία και μόνο προσαρμογή του μοντέλου στο σύνολο των n παρατηρήσεων ως εξής:

$$\text{MSE}_{\text{PRESS}}(h_x) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \widehat{y}(x_i)}{1 - S_{ii}(h_x)} \right)^2.$$

Στον παρακάτω κώδικα εφαρμόζεται η εκτίμηση του σφάλματος Leave-One-Out Cross-Validation για τον εκτιμητή Nadaraya-Watson, μέσω του τύπου PRESS (Predicted Residual Sum of Squares). Για κάθε τιμή του bandwidth h_x, κατασκευάζεται ο πίνακας εξομάλυνσης (smoother matrix) S, ο οποίος περιέχει τα σταθμισμένα βάρη Gaussian πυρήνα για κάθε σημείο του δείγματος. Οι εκτιμήσεις $\hat{y}_i$ προκύπτουν από τον γραμμικό συνδυασμό $S \cdot y$, ενώ τα διαγώνια στοιχεία S_{ii} εκφράζουν το self-influence (βαθμό αυτο-επιρροής) κάθε παρατήρησης. Στη συνέχεια, εκτιμάται το σφάλμα LOOCV μέσω του τύπου PRESS, ο οποίος διορθώνει το σφάλμα κάθε παρατήρησης λαμβάνοντας υπόψη τη δική της συμβολή στην προσαρμοσμένη τιμή μέσω του S_{ii}. Έτσι, ο επαναληπτικός υπολογισμός n μοντέλων αντικαθίσταται από πράξεις πινάκων, μειώνοντας σημαντικά το υπολογιστικό κόστος. Η διαδικασία αυτή επαναλαμβάνεται για όλες τις υποψήφιες τιμές του h_x, και επιλέγεται τελικά εκείνη που ελαχιστοποιεί το PRESS error, προσφέροντας έτσι μια υπολογιστικά αποδοτικότερη και αριθμητικά ισοδύναμη εκτίμηση του LOOCV σφάλματος.

```r
# 1. Read the data
data <- readRDS("data1pairs.rds")
x <- data[, 1]
y <- data[, 2]
n <- length(x)

# 2. Define grid of bandwidth values
h_vals <- seq(0.05, 1, by = 0.01)
press_errors <- numeric(length(h_vals))  # Store PRESS errors for each h

# 3. Efficient PRESS computation for each h
for (j in seq_along(h_vals)) {
  h <- h_vals[j]

  # Construct the Smoother Matrix S (n x n)
  S <- matrix(0, nrow = n, ncol = n)

  for (i in 1:n) {
    # Compute Gaussian kernel weights for the i-th observation
    weights <- dnorm((x[i] - x) / h)
    # Normalize weights to sum to 1 (row of smoother matrix S)
    S[i, ] <- weights / sum(weights)
  }

  # Compute fitted values: y_hat = S * y
  y_hat <- S %*% y

  # Extract diagonal elements S_ii (self-influence)
  S_ii <- diag(S)

  # Apply PRESS formula
  residuals <- y - y_hat
  press_residuals <- residuals / (1 - S_ii)

  # Compute mean squared PRESS error for current h
  press_errors[j] <- mean(press_residuals^2)
}

# 4. Find optimal h minimizing PRESS error
best_h <- h_vals[which.min(press_errors)]
cat("Optimal bandwidth h:", best_h, "\n")
```

Κατά την εφαρμογή της μη παραμετρικής παλινδρόμησης, η εύρεση του βέλτιστου πλάτους h_x πραγματοποιήθηκε με δύο διαφορετικές, πλην θεωρητικά ισοδύναμες, προσεγγίσεις. Η πρώτη προσέγγιση, που βασίστηκε στην επαναληπτική χρήση της συνάρτησης `ksmooth` εντός ενός βρόχου Leave-One-Out Cross-Validation (LOOCV), υπέδειξε ως βέλτιστη τιμή $h_x \approx 0.86$. Αντίθετα, η δεύτερη, υπολογιστικά αποδοτικότερη προσέγγιση που αξιοποιεί τον αναλυτικό τύπο του PRESS, κατέληξε σε μια σημαντικά διαφορετική βέλτιστη τιμή, $h_x \approx 0.32$.

Σε αυτό το σημείο κρίνεται σκόπιμο να οπτικοποιηθεί η καμπύλη Nadaraya-Watson καθώς και η καμπύλη του MSE, προκειμένου να ελεγχθεί η ορθότητα των αποτελεσμάτων και να γίνει σύγκριση της προσαρμογής στα δεδομένα. Ο παρακάτω κώδικας υλοποιεί τα εν λόγω διαγράμματα.

```r
# Plot PRESS criterion
df_press <- data.frame(h = h_vals, press_mse = press_errors)

ggplot(df_press, aes(x = h, y = press_mse)) +
  geom_line(color = "brown4", linewidth = 1.2) +
  geom_vline(xintercept = best_h, color = "darksalmon", linetype = "dashed",
    linewidth = 1) +
  labs(
    title = "Bandwidth Selection via PRESS Criterion for Nadaraya-Watson",
    subtitle = paste("Optimal bandwidth h =", round(best_h, 3)),
    x = "Bandwidth (h)",
    y = "PRESS Error (MSE)",
    caption = "Efficient PRESS computation using smoother matrix S") +
  theme_gray(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5, margin = margin(b = 10)),
    axis.title = element_text(face = "bold"),
    plot.caption = element_text(size = 10, face = "italic", hjust = 1))

# Prepare dataframes for plotting Nadaraya-Watson regression estimate
df_points <- data.frame(x = x, y = y)

# Compute fitted values directly on observed x using optimal h
h_opt <- best_h
y_hat <- numeric(n)

for (i in 1:n) {
  weights <- dnorm((x[i] - x) / h_opt)
  y_hat[i] <- sum(weights * y) / sum(weights)}

df_fit <- data.frame(x = x, y_hat = y_hat)

# Plot Nadaraya-Watson regression estimate
ggplot() +
  geom_point(data = df_points, aes(x = x, y = y), color = "darksalmon", size = 2.5,
     alpha = 0.7) +
  geom_line(data = df_fit, aes(x=x, y=y_hat), color = "brown4", linewidth = 1.2)+
  labs(
    title = "Nadaraya-Watson Regression Estimate (Gaussian kernel)",
    subtitle = paste("Optimal bandwidth h =", round(h_opt, 3), "(from PRESS
    criterion)"),
    x = "x",
    y = "y",
    caption = "Efficient bandwidth selection using PRESS formula") +
  theme_gray(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5, margin = margin(b = 10)),
    axis.title = element_text(face = "bold"),
    plot.caption = element_text(size = 10, face = "italic", hjust = 1))
```

Στο Σχήμα 2 παρουσιάζει την καμπύλη του Μέσου Τετραγωνικού Σφάλματος (MSE), όπως αυτό υπολογίστηκε μέσω του τύπου PRESS, για ένα εύρος υποψήφιων τιμών του h_x. Το ελάχιστο της καμπύλης, που υποδεικνύεται από την κόκκινη διακεκομμένη γραμμή, αντιστοιχεί στη βέλτιστη τιμή του πλάτους, η οποία βρέθηκε ίση με $h_x = 0.32$.

Το Σχήμα 3 απεικονίζει την τελική εκτίμηση της συνάρτησης παλινδρόμησης με τη μέθοδο Nadaraya-Watson. Η συνεχής κόκκινη καμπύλη, η οποία υπολογίστηκε χρησιμοποιώντας το βέλτιστο bandwidth $h_x = 0.32$ (που προέκυψε από το κριτήριο PRESS), προσαρμόζεται ομαλά στα δεδομένα, συλλαμβάνοντας επιτυχώς τη μη-γραμμική δομή τους χωρίς να υπερπροσαρμόζεται στον θόρυβο.

Η φαινομενική αντίφαση όσον αφορά το βέλτιστο bandwidth που προέκυψε από την άμεση χρήση της συνάρτησης `ksmooth` με LOOCV ($h_x \approx 0.86$), σε αντιδιαστολή με την τιμή που υπολογίστηκε με τη μέθοδο PRESS ($h_x \approx 0.32$) δεν υποδηλώνει σφάλμα σε καμία από τις δύο μεθόδους. Πιθανότατα πηγάζει από μια λεπτή διαφορά στον ορισμό της παραμέτρου bandwidth στην υλοποίηση της συνάρτησης `ksmooth`. Μαθηματικά, ο εκτιμητής Nadaraya-Watson ορίζεται από τον πυρήνα $K(u)$ και το πλάτος h_x. Στην υλοποίησή με τον τύπο PRESS, το h_x αντιστοιχεί απευθείας στην τυπική απόκλιση του Gaussian πυρήνα που χρησιμοποιήθηκε. Ωστόσο, η τεκμηρίωση (documentation) της `ksmooth` αποκαλύπτει ότι η παράμετρος bandwidth που δέχεται η συνάρτηση δεν είναι το ίδιο το h_x, αλλά μια τιμή που εσωτερικά επανακλιμακώνεται [4]. Συνεπώς, και οι δύο προσεγγίσεις φαίνεται να κατέληξαν σε παραπλήσιο συμπέρασμα για το βέλτιστο πλάτος h_x.

Ερώτημα 1 (α) iii: Συμπεριφορά της Εκτίμησης Nadaraya-Watson σε Οριακές Τιμές h_x

Η μέθοδος Nadaraya-Watson εκτιμά τη συνάρτηση παλινδρόμησης $m(x)$ ως σταθμισμένο μέσο όρο των παρατηρούμενων τιμών y_i, με βάρη που εξαρτώνται από την απόσταση των x_i από το σημείο x και το bandwidth h_x. Το μέγεθος του bandwidth καθορίζει το εύρος της περιοχής γύρω από κάθε x που λαμβάνεται υπόψη κατά την εκτίμηση και ρυθμίζει την ισορροπία μεταξύ μεροληψίας και διακύμανσης (bias-variance tradeoff).

Στην περίπτωση $h_x \to 0$ (πολύ μικρό bandwidth), ο πυρήνας $K\left(\frac{x-x_i}{h}\right)$ συγκεντρώνει σχεδόν όλο το βάρος σε παρατηρήσεις που βρίσκονται εξαιρετικά κοντά στο σημείο x.Στο όριο $h \to 0$, για κάθε x, ο εκτιμητής τείνει να λαμβάνει υπόψη μόνο την τιμή y_i της πλησιέστερης παρατήρησης $x_i \approx x$. Συνεπώς, η εκτίμηση καταλήγει:

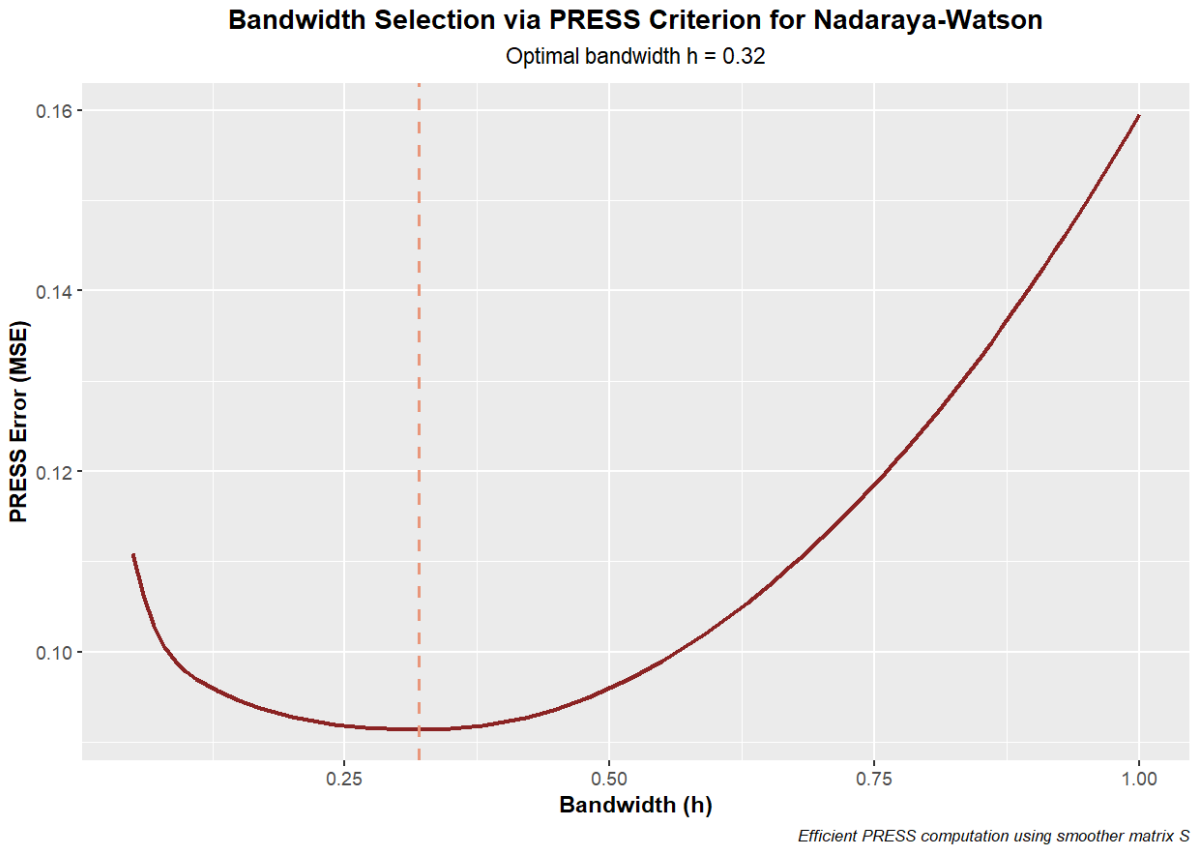
$$\hat{m}(x) \approx y_i \quad \text{όπου} \quad x_i \approx x.$$

Δηλαδή, η καμπύλη τείνει να περνάει σχεδόν από όλα τα σημεία του δείγματος. Η συμπεριφορά αυτή αντιστοιχεί στην κλασική περίπτωση της υπερπροσαρμογής (overfitting). Το μοντέλο έχει σχεδόν μηδενική μεροληψία (low bias), καθότι δεν κάνει απλοϊκές παραδοχές και προσπαθεί να αποτυπώσει κάθε λεπτομέρεια των δεδομένων εκπαίδευσης, και εξαιρετικά υψηλή διακύμανση (high variance), καθώς είναι υπερβολικά ευαίσθητο στις μικρές διακυμάνσεις του δείγματος και αποτυγχάνει να γενικεύσει σε νέα δεδομένα.

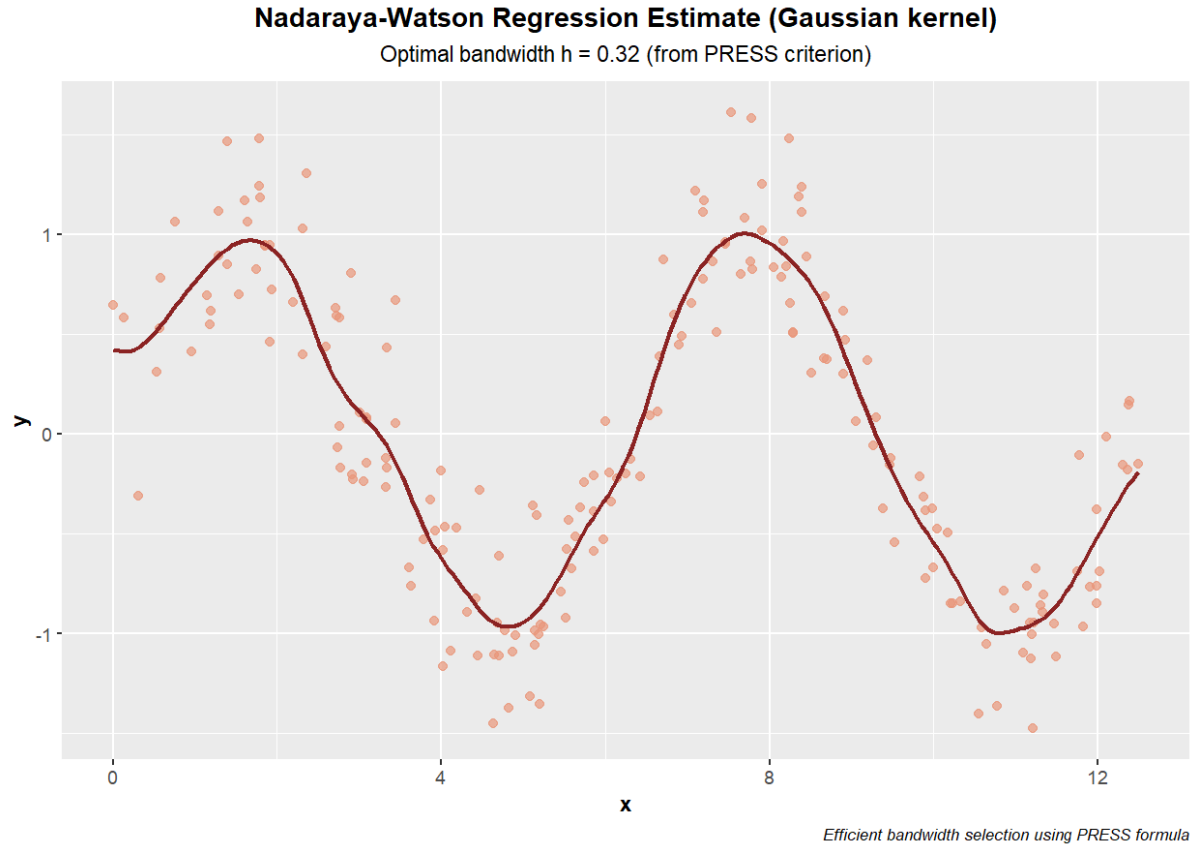
Στην περίπτωση $h_x \to \infty$ (πολύ μεγάλο bandwidth), ο πυρήνας γίνεται σχεδόν επίπεδος και εκχωρεί σχεδόν ίσο βάρος σε όλες τις παρατηρήσεις, ανεξάρτητα από την τιμή του x. Δηλαδή $K\left(\frac{x-x_i}{h}\right) \approx$ σταθερό $\quad \forall i$. Κατά συνέπεια, ο εκτιμητής τείνει να υπολογίζει τον συνολικό μέσο όρο των y_i:

$$\hat{m}(x) \approx \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Η εκτιμώμενη καμπύλη είναι πρακτικά μια οριζόντια ευθεία, πλήρως εξομαλυμένη και αγνοεί κάθε τοπική πληροφορία που παρέχεται από τη σχέση μεταξύ X και Y. Αυτή η συμπεριφορά αντιστοιχεί σε υποπροσαρμογή (underfitting), όπου το μοντέλο έχει πολύ υψηλή μεροληψία, καθώς βασίζεται στην υπερβολικά απλοϊκή παραδοχή ότι η συνάρτηση είναι σταθερή, και σχεδόν μηδενική διακύμανση, αφού είναι σταθερό και δεν επηρεάζεται από μεμονωμένα σημεία.



Σχήμα 2. *Επιλογή του βέλτιστου bandwidth για τον εκτιμητή Nadaraya-Watson μέσω του κριτηρίου PRESS (μέσο τετραγωνικό σφάλμα).*



Σχήμα 3. *Εκτίμηση συνάρτησης παλινδρόμησης με τη μέθοδο Nadaraya-Watson και βέλτιστο bandwidth* $h = 0.86$ *(Gaussian πυρήνας, επιλογή μέσω Leave-One-Out Cross-Validation).*

Ερώτημα 1 (β) i: Μελέτη της Κατανομής $T = \min(X_1, X_2, \ldots, X_n)$ με τη Μέθοδο Bootstrap

Μελετάται η κατανομή της στατιστικής συνάρτησης $T = \min(X_1, X_2, \ldots, X_n)$ μέσω της μεθόδου Bootstrap για τυχαίο δείγμα μεγέθους $n = 200$ (`data1b.rds`) από μια άγνωστη συνεχή κατανομή.

Η μέθοδος Bootstrap είναι μια τεχνική επαναδειγματοληψίας (resampling) που επιτρέπει την προσέγγιση της κατανομής δειγματοληψίας μιας στατιστικής συνάρτησης, όταν η πραγματική κατανομή του πληθυσμού είναι άγνωστη [5]. Η κεντρική ιδέα βασίζεται στην αρχή της αντικατάστασης (plug-in principle). Δεδομένου ότι δεν έχουμε πρόσβαση στην πραγματική συνάρτηση κατανομής του πληθυσμού F, την αντικαθιστούμε με την καλύτερη διαθέσιμη εκτίμησή της, την εμπειρική συνάρτηση κατανομής (Empirical Distribution Function - EDF), $\widehat{E}_n$, που προκύπτει από το αρχικό μας δείγμα. Η διαδικασία που ακολουθείται είναι η εξής:

1. Από το αρχικό δείγμα $x = (x_1, \ldots, x_n)$, υπολογίζεται η παρατηρηθείσα τιμή της στατιστικής συνάρτησης, $T_{\text{obs}} = \min(x)$.

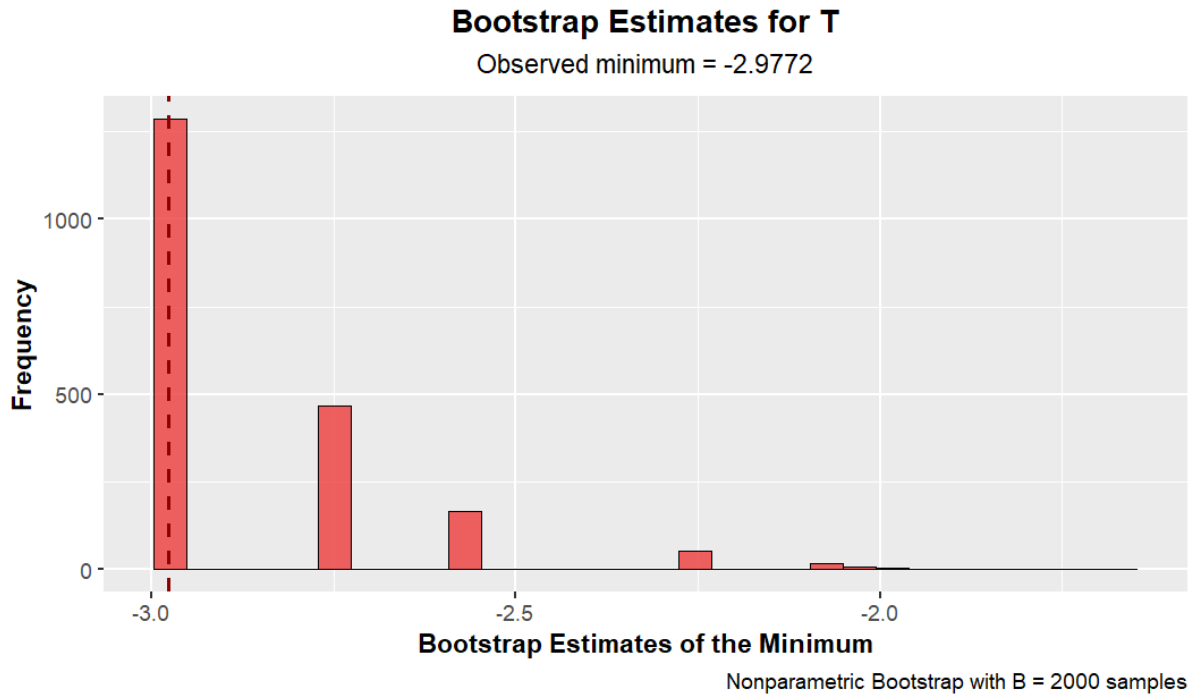
2. Δημιουργούνται B νέα bootstrap δείγματα, $x^{*1}, x^{*2}, \ldots, x^{*B}$, το καθένα μεγέθους n. Κάθε bootstrap δείγμα προκύπτει μέσω δειγματοληψίας με επανατοποθέτηση από το αρχικό δείγμα x.

3. Για κάθε bootstrap δείγμα x^{*b}, υπολογίζεται η αντίστοιχη bootstrap εκτίμηση της στατιστικής συνάρτησης, $T^{*b} = \min(x^{*b})$.

4. Η συλλογή των B τιμών $(T^{*1}, \ldots, T^{*B})$ αποτελεί μια εμπειρική προσέγγιση της άγνωστης κατανομής δειγματοληψίας της T. Το ιστόγραμμα αυτών των τιμών δίνει μια οπτικοποίηση αυτής της προσέγγισης.

Η διαδικασία αυτή υλοποιείται με τον παρακάτω κώδικα. Αρχικά, φορτώνεται το δείγμα από το αρχείο `data1b.rds` και υπολογίζεται η παρατηρηθείσα ελάχιστη τιμή `T_original`. Στη συνέχεια, ορίζονται οι παράμετροι της μεθόδου ($B = 2000$ bootstrap δείγματα) και εκτελείται ένας βρόχος for loop B φορές. Σε κάθε επανάληψη δημιουργείται ένα νέο bootstrap δείγμα μεγέθους n με τη συνάρτηση `sample`, η οποία πραγματοποιεί δειγματοληψία με επανατοποθέτηση από τα αρχικά δεδομένα και υπολογίζεται η ελάχιστη τιμή του bootstrap δείγματος και αποθηκεύεται στο διάνυσμα `bootstrap_mins`. Τέλος, κατασκευάζεται το ιστόγραμμα των 2000 bootstrap εκτιμήσεων και προστίθεται μια διακεκομμένη κάθετη γραμμή που αντιστοιχεί στην ελάχιστη τιμή του αρχικού δείγματος για λόγους σύγκρισης.

```r
# Load data
data <- readRDS("data1b.rds")
n <- length(data)

# Compute the observed minimum
T_original <- min(data)

# Set bootstrap parameters
B <- 2000
set.seed(123)

# Compute bootstrap estimates of the minimum
bootstrap_mins <- numeric(B)
for (b in 1:B) {
  sample_b <- sample(data, size = n, replace = TRUE)
  bootstrap_mins[b] <- min(sample_b)
}

df_boot <- data.frame(min_values = bootstrap_mins)
```



Σχήμα 4. *Ιστόγραμμα των $B = 2000$ bootstrap εκτιμήσεων της συνάρτησης $T = \min(X_1, \ldots, X_n)$ με παρατηρηθέν ελάχιστο $T_{obs} = -2.9772$.*

```r
# Plot histogram using ggplot2
ggplot(df_boot, aes(x = min_values)) +
  geom_histogram(fill = "brown2", color = "black", bins = 30, alpha = 0.8) +
  geom_vline(xintercept = T_original, color = "darkred", linetype = "dashed",
    linewidth = 1) +
  labs(
    title = "Bootstrap Estimates for T",
    subtitle = paste("Observed minimum =", round(T_original, 4)),
    x = "Bootstrap Estimates of the Minimum",
    y = "Frequency",
    caption = "Nonparametric Bootstrap with B = 2000 samples" ) +
  theme_gray(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5, margin = margin(b = 10)),
    axis.title = element_text(face = "bold"),
    legend.position = "none")
```

Το ιστόγραμμα που προκύπτει από την εφαρμογή της μεθόδου Bootstrap παρουσιάζεται στο Σχήμα 4. Παρατηρείται ότι η κατανομή των bootstrap εκτιμήσεων είναι έντονα ασύμμετρη προς τα δεξιά, με τη συντριπτική πλειοψηφία των εκτιμήσεων να είναι ίση με την ελάχιστη τιμή του αρχικού δείγματος, $T_{obs} \approx -2.9772$. Επίσης, δεν υπάρχει καμία bootstrap εκτίμηση που να είναι μικρότερη από την τιμή αυτή.

Η μέθοδος Bootstrap αποτυγχάνει σε αυτή την περίπτωση, διότι η στατιστική συνάρτηση του ελαχίστου είναι μια ακραία στατιστική συνάρτηση (extreme statistic), η οποία δεν είναι λεία (non-smooth). Ειδικότερα, η διαδικασία της δειγματοληψίας με επανατοποθέτηση σημαίνει ότι κάθε στοιχείο ενός bootstrap δείγματος x^* προέρχεται από το αρχικό δείγμα x. Κατά συνέπεια, κάθε τιμή στο x^* είναι αναγκαστικά μεγαλύτερη ή ίση με την ελάχιστη τιμή του αρχικού δείγματος (T_{obs}). Επομένως, η ελάχιστη τιμή οποιουδήποτε bootstrap δείγματος,

T^*, δεν μπορεί ποτέ να είναι μικρότερη από την T_{obs}: $min(x^{*b} \geq min(x) \implies T^* \geq T_{obs}$. Η κατανομή που παράγεται είναι εξ ορισμού censored στην τιμή T_{obs}, γεγονός που οδηγεί σε μια εξαιρετικά μεροληπτική (biased) εκτίμηση της πραγματικής κατανομής δειγματοληψίας. Φαίνεται λοιπόν ότι η μέθοδος Bootstrap δεν είναι κατάλληλη για την εκτίμηση της κατανομής δειγματοληψίας ακραίων τιμών, όπως το ελάχιστο, διότι η φύση της επαναδειγματοληψίας περιορίζει τεχνητά το εύρος των εκτιμήσεων και εισάγει ισχυρή μεροληψία [3].

Ερώτημα 1 (β) ii: Μελέτη της Κατανομής $T = \min(X_1, \ldots, X_n)$ με Παραμετρικό Bootstrap

Στο προηγούμενο ερώτημα, διαπιστώθηκε ότι η μη-παραμετρική μέθοδος Bootstrap αποτυγχάνει να εκτιμήσει την κατανομή δειγματοληψίας της ελάχιστης τιμής, λόγω της φύσης της επαναδειγματοληψίας που περιορίζει τις εκτιμήσεις να είναι πάντα μεγαλύτερες ή ίσες με την παρατηρηθείσα ελάχιστη τιμή. Έτσι, εξετάζεται η περίπτωση όπου έχουμε την επιπλέον πληροφορία ότι το δείγμα προέρχεται από μια κατανομή Student (t-distribution). Η διαδικασία διαφοροποιείται από τη μη-παραμετρική προσέγγιση στο βήμα της παραγωγής των bootstrap δειγμάτων. Αντί να γίνεται δειγματοληψία από την εμπειρική συνάρτηση κατανομής (EDF), η διαδικασία έχει ως εξής:

1. Υποθέτουμε ότι τα δεδομένα $x = (x_1, \ldots, x_n)$ προέρχονται από μια κατανομή $F(\theta)$ που ανήκει σε μια γνωστή οικογένεια, εν προκειμένω την κατανομή Student-t. Οι άγνωστες παράμετροι θ αυτής της κατανομής είναι οι βαθμοί ελευθερίας (df), η μέση τιμή (μ) και η τυπική απόκλιση (σ).

2. Χρησιμοποιώντας το αρχικό δείγμα x, γίνεται εκτίμηση των άγνωστων παραμέτρων, λαμβάνοντας τις εκτιμήσεις $\hat{\theta} = (df, \mu, \sigma)$.

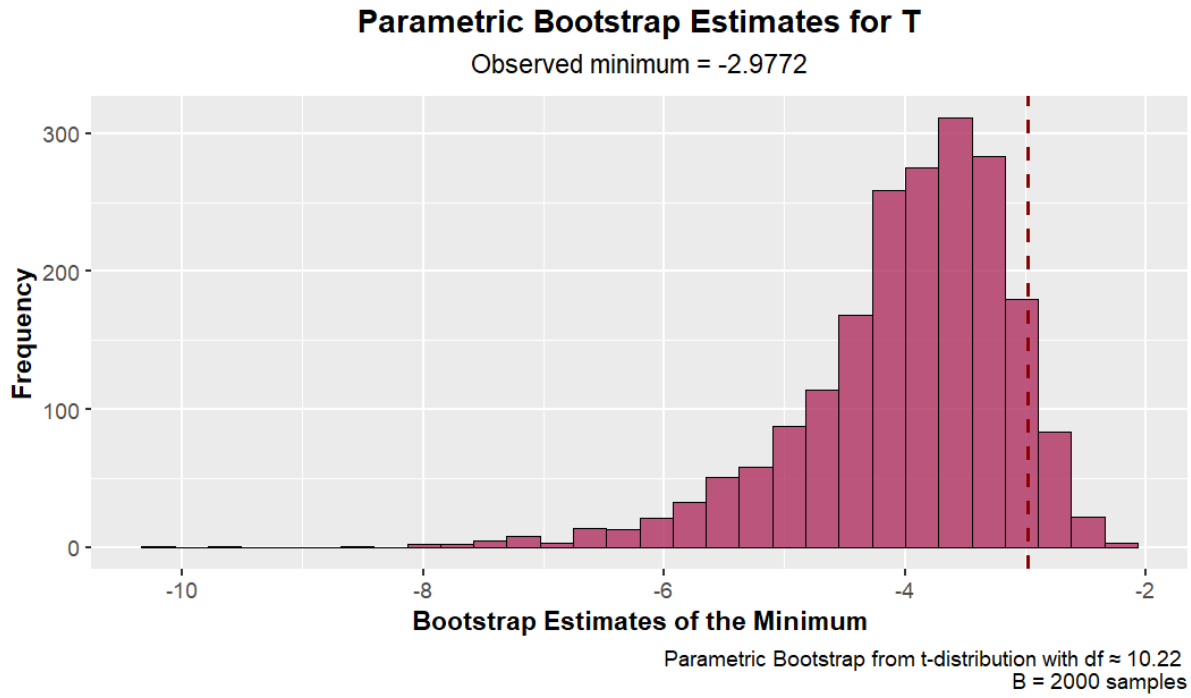
3. Παράγονται B νέα bootstrap δείγματα, $x^{*1}, \ldots, x^{*B}$, μεγέθους n, αντλώντας τιμές από την προσαρμοσμένη παραμετρική κατανομή $F(\hat{\theta})$.

4. Για κάθε bootstrap δείγμα x^{*b}, υπολογίζεται η εκτίμηση $T^{*b} = \min(x^{*b})$.

Το σύνολο των B τιμών T^* αποτελεί την παραμετρική bootstrap προσέγγιση της κατανομής δειγματοληψίας της T.

Ο κώδικας που αναπτύχθηκε για την υλοποίηση του παραμετρικού bootstrap ακολουθεί τα παραπάνω βήματα. Αρχικά, εκτιμώνται οι παράμετροι της κατανομής Student-t από τα δεδομένα. Η μέση τιμή (μ) εκτιμάται από τον δειγματικό μέσο (`mean(data)`), η τυπική απόκλιση (σ) από τη δειγματική τυπική απόκλιση (`sd(data)`), και οι βαθμοί ελευθερίας df εκτιμώνται με τη μέθοδο της μέγιστης πιθανοφάνειας, χρησιμοποιώντας τη συνάρτηση `fitdistr()` από τη βιβλιοθήκη `MASS`. Για την παραγωγή των Bootstrap δειγμάτων εκτελείται $B = 2000$ φορές ένας βρόχος. Σε κάθε επανάληψη παράγεται ένα νέο δείγμα μεγέθους n από την τυπική κατανομή Student-t με τους εκτιμηθέντες βαθμούς ελευθερίας (`rt(n, df = df_hat)`). Το δείγμα αυτό μετασχηματίζεται (κλιμακώνεται και μετατοπίζεται) ώστε να αντιστοιχεί στις εκτιμηθείσες παραμέτρους $\widehat{\mu}$ και $\widehat{\sigma}$. Υπολογίζεται η ελάχιστη τιμή του νέου αυτού δείγματος και αποθηκεύεται. Τέλος, κατασκευάζεται το ιστόγραμμα των 2000 παραμετρικών bootstrap εκτιμήσεων με τη βιβλιοθήκη `ggplot2`.

```r
# Load data
data <- readRDS("data1b.rds")
n <- length(data)

# Compute the observed minimum
T_original <- min(data)

# Estimate parameters for the Student's t-distribution
mu_hat <- mean(data)
```



Σχήμα 5. *Ιστόγραμμα $B = 2000$ παραμετρικών bootstrap εκτιμήσεων της συνάρτησης $T = \min(X_1, \ldots, X_n)$ με παρατηρηθέν ελάχιστο $T_{obs} = -2.9772$. Θεωρείται ότι το δείγμα έχει προέλθει από κατανομή Student με βαθμούς ελευθερίας $df = 10.22$.*

```r
sd_hat <- sd(data)
df_hat <- fitdistr(data, densfun = "t")$estimate["df"]

# Set bootstrap parameters
B <- 2000
set.seed(123)

# Generate parametric bootstrap estimates for the minimum
parametric_mins <- numeric(B)
for (b in 1:B) {
  # Generate sample from t-distribution with estimated parameters
  sample_b <- rt(n, df = df_hat) * sd_hat + mu_hat
  parametric_mins[b] <- min(sample_b)
}

# Plot histogram using ggplot2
df_parametric <- data.frame(min_values = parametric_mins)

ggplot(df_parametric, aes(x = min_values)) +
  geom_histogram(fill = "maroon", color = "black", bins = 30, alpha = 0.8) +
  geom_vline(xintercept = T_original, color = "darkred", linetype = "dashed",
    linewidth = 1) +
  labs(
    title = "Parametric Bootstrap Estimates for T",
    subtitle = paste("Observed minimum =", round(T_original, 4)),
    x = "Bootstrap Estimates of the Minimum",
```

```
    y = "Frequency",
    caption = paste("Parametric Bootstrap from t-distribution with df ≈", round(df_
    hat, 2), "\nB = 2000 samples")
  ) +
  theme_gray(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5, margin = margin(b = 10)),
    axis.title = element_text(face = "bold"),
    legend.position = "none")
```

Το ιστόγραμμα που προκύπτει από την εφαρμογή της παραμετρικής μεθόδου Bootstrap παρουσιάζεται στο Σχήμα 5. Με την παραμετρική μέθοδο Bootstrap, η κατανομή των εκτιμήσεων της $T = \min(X_1, \ldots, X_n)$ εμφανίζει πλέον μια ομαλή και συνεχή μορφή, χωρίς τη συσσώρευση στο παρατηρηθέν ελάχιστο που παρατηρήθηκε στην μη-παραμετρική εκδοχή. Επιπλέον, η θεωρητική t-κατανομή, από την οποία προέρχονται τα bootstrap δείγματα, επιτρέπει την εμφάνιση τιμών σε ένα ευρύ φάσμα, με αποτέλεσμα να παρατηρούνται και τιμές μικρότερες από το $T_{\text{obs}} = -2.9772$, κάτι που δεν ήταν εφικτό προηγουμένως. Ακόμη, το T_{obs} δεν αποτελεί πλέον ακραία τιμή της κατανομής αλλά εντάσσεται στο κυρίως σώμα της και άρα μπορεί να θεωρηθεί ως μια σχετικά τυπική εκτίμηση της T, υπό την υπόθεση ότι τα δεδομένα ακολουθούν t-κατανομή με παραμέτρους προσαρμοσμένους από το δείγμα.

Συμπερασματικά, η παραμετρική προσέγγιση επιτυγχάνει μια πιο ρεαλιστική απεικόνιση της κατανομής δειγματοληψίας της T, αρκεί η υπόθεση για την κατανομή του πληθυσμού να είναι εύλογη. Σε αντίθεση με τη μη-παραμετρική μέθοδο, περιορίζεται η επίδραση μεμονωμένων ακραίων παρατηρήσεων, προσφέροντας έτσι πιο σταθερές και "λογικές" εκτιμήσεις.

Άσκηση 2

Ερώτημα 2 (α) i: Προσομοίωση Τιμών από την Τυποποιημένη Κανονική Κατανομή, $\mathcal{N}(0,1)$, με τη Μέθοδο "Squeezed Rejection Sampling"

Ζητείται η παραγωγή ενός τυχαίου δείγματος 10,000 τιμών από την τυπική κανονική κατανομή, $f(x) = \mathcal{N}(0,1)$ με τη μέθοδο. Επειδή η συνάρτηση αθροιστικής πιθανότητας της κανονικής κατανομής δεν αντιστρέφεται αναλυτικά, δε μπορεί να εφαρμοστεί η μέθοδος της αντιστροφής. Έτσι, χρησιμοποιείται μια παραλλαγή της μεθόδου της απόρριψης, η Squeezed Rejection Sampling.

Η μέθοδος της απόρριψης (Rejection Sampling) είναι μια τεχνική προσομοίωσης από μια κατανομή-στόχο $f(x)$ (target density) και βασίζεται στη χρήση μιας απλούστερης "κατανομής εισήγησης" $g(x)$ (proposal density), από την οποία μπορούμε εύκολα να παράγουμε δείγματα. Εισάγεται επίσης μια σταθερά M, τέτοια ώστε $f(x) \leq M \cdot g(x)$ για κάθε x για το οποίο ισχύει $f(x) > 0$. Η συνάρτηση $M \cdot g(x)$ ονομάζεται συνάρτηση «φακέλου» (envelope).

Η Squeezed Rejection Sampling αποτελεί βελτίωση της παραπάνω διαδικασίας, με στόχο τη μείωση του υπολογιστικού κόστους, ειδικά όταν ο υπολογισμός της $f(x)$ είναι χρονοβόρος. Εισάγεται μια τρίτη συνάρτηση $s(x)$, που ονομάζεται συνάρτηση συμπίεσης (squeezing function), η οποία είναι εύκολη στον υπολογισμό και για κάθε x για το οποίο ισχύει $f(x) > 0$ ικανοποιεί τη συνθήκη [6, 7]

$$s(x) \leq f(x) \leq M \cdot g(x).$$

Ο αλγόριθμος περιλαμβάνει αρχικά την παραγωγή ενός υποψήφιου δείγματος $y \sim g(x)$ και μιας τυχαίας τιμής $u \sim \mathcal{U}(0,1)$. Κατά τον γρήγορο έλεγχο (Squeeze Test), ελέγχεται πρώτα αν $u \leq \frac{s(y)}{M \cdot g(y)}$, όπου u είναι μια τυχαία τιμή από την $\mathcal{U}(0,1)$. Αν η συνθήκη ισχύει, τότε το y γίνεται απευθείας αποδεκτό και δεν χρειάζεται ο υπολογισμός της $f(y)$. Αυτό είναι το βασικό πλεονέκτημα της μεθόδου. Αν ο παραπάνω γρήγορος έλεγχος αποτύχει, τότε και μόνο τότε προχωράμε στον πλήρη έλεγχο (Full Test). Σε αυτή την περίπτωση, υπολογίζεται η $f(y)$ και εκτελείται ο κανονικός έλεγχος της μεθόδου απόρριψης: $u \leq \frac{f(y)}{M \cdot g(y)}$. Αν ισχύει, το y γίνεται αποδεκτό, αλλιώς απορρίπτεται.

Στην παρούσα εργασία, ως κατανομή εισήγησης (proposal density) χρησιμοποιείται η κατανομή Laplace,

$$g(x) = \frac{1}{2} e^{-|x|},$$

η σταθερά M δίνεται από τη σχέση $M = \sqrt{\frac{2e}{\pi}}$, και η συνάρτηση συμπίεσης (squeezing function) $s(x)$ προκύπτει από την ανισότητα $e^{-x^2/2} \geq 1 - \frac{x^2}{2}$, δηλαδή:

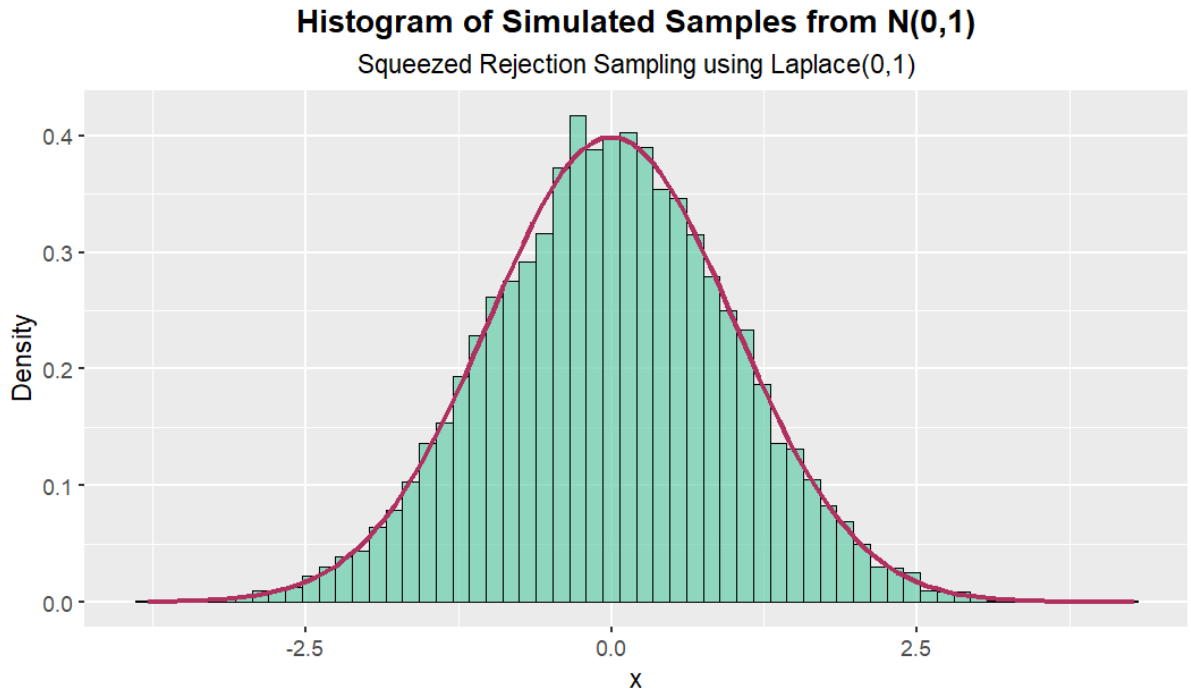
$$s(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} \left(1 - \dfrac{x^2}{2}\right), & \text{όταν } |x| < \sqrt{2}, \\ 0, & \text{διαφορετικά.} \end{cases},$$

η οποία είναι υπολογιστικά πολύ φθηνότερη από την $f(x)$ καθώς δεν περιλαμβάνει εκθετικές συναρτήσεις. Η παραγωγή δειγμάτων από την $g(x)$ υλοποιείται με τη μέθοδο της αντιστροφής, αξιοποιώντας την αναλυτική μορφή της αντίστροφης αθροιστικής συνάρτησης πιθανότητας της κατανομής Laplace.

Ο ακόλουθος κώδικας υλοποιεί την παραπάνω μεθοδολογία. Αρχικα, ορίζονται οι συναρτήσεις πυκνότητας για την κατανομή στόχο $f(x)$ (`dnorm(x, 0, 1)`), την κατανομή εισήγησης $g(x)$ Laplace (g) και την κατανομή συμπίεσης, `squeezing(x)`. Δημιουργείται η συνάρτηση `sample_laplace` για την παραγωγή δειγμάτων από την κατανομή Laplace μέσω της μεθόδου της αντιστροφής της CDF της. Η βασική λειτουργία υλοποιείται μέσω ενός βρόχου while, ο οποίος συνεχίζει να εκτελείται μέχρι να συγκεντρωθεί ο επιθυμητός αριθμός των `n_samples`. Μέσα στον βρόχο, πραγματοποιούνται οι έλεγχοι αποδοχής (πρώτα ο "squeeze test"

και μετά, αν χρειαστεί, ο πλήρης έλεγχος απόρριψης) και μετράται ο αριθμός των αποδοχών σε κάθε στάδιο. Τέλος κατασκευάζεται ένα ιστόγραμμα πυκνότητας των 10.000 προσομοιωμένων τιμών, καθώς επίσης και η θεωρητική καμπύλη της $\mathcal{N}(0,1)$ για σύγκριση.

```r
# Define target density: Standard Normal N(0,1)
f <- function(x) dnorm(x)

# Define proposal density: Laplace(0,1)
g <- function(x) 0.5 * exp(-abs(x))

# Inverse CDF sampling from Laplace(0,1)
sample_laplace <- function(n) {
  u <- runif(n)
  ifelse(u < 0.5, log(2 * u), -log(2 * (1 - u)))
}

# Envelope constant M
M <- sqrt(2 * exp(1) / pi)

# Squeezing function: lower bound for f(x)
squeezing <- function(x) {
  val <- 1 - (x^2 / 2)
  ifelse(val > 0, val / sqrt(2 * pi), 0)}

# Number of samples to generate
n_samples <- 10000
samples <- numeric(n_samples)
accepted <- 0
set.seed(42)

# Counters for acceptance rates
total <- squeeze_accepts <- full_accepts <- 0

# Squeezed rejection sampling loop
while (accepted < n_samples) {
  y <- sample_laplace(1)
  u <- runif(1)
  total <- total + 1

  g_y <- g(y)
  s_y <- squeezing(y)

  # First test: squeezing acceptance
  if (u <= s_y / (M * g_y)) {
    accepted <- accepted + 1
    samples[accepted] <- y
    squeeze_accepts <- squeeze_accepts + 1
  }
  # Second test: full evaluation of f(x)
  else if (u <= f(y) / (M * g_y)) {
    accepted <- accepted + 1
```



Σχήμα 6. *Ιστόγραμμα* 10000 *τιμών από την* $N(0,1)$ *που προσομοιώθηκαν με τη μέθοδο squeezed rejection sampling, χρησιμοποιώντας ως εισήγηση την κατανομή Laplace(0,1) και σταθερά* $M = \sqrt{\frac{2e}{\pi}}$. *Η υπέρθεση της θεωρητικής πυκνότητας* $N(0,1)$ *επιβεβαιώνει την ακρίβεια της προσομοίωσης.*

```
    samples[accepted] <- y
    full_accepts <- full_accepts + 1}}

# Plot histogram with theoretical normal curve
df <- data.frame(x = samples)

ggplot(df, aes(x = x)) +
  geom_histogram(aes(y = ..density..), bins = 60, fill = "aquamarine3", color = "
    black", alpha = 0.7) +
  stat_function(fun = dnorm, color = "maroon", size = 1.2) +
  labs(title = "Histogram of Simulated Samples from N(0,1)",
       subtitle = "Squeezed Rejection Sampling using Laplace(0,1)",
       x = "x", y = "Density") +
  theme_gray(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

# Print acceptance statistics
cat("Total proposals:", total, "\n")
cat("Accepted via squeezing:", squeeze_accepts, "\n")
cat("Accepted via full f(x):", full_accepts, "\n")
```

Τα αποτελέσματα της προσομοίωσης παρουσιάζονται στο Σχήμα 6. Είναι εμφανές ότι το ιστόγραμμα των 10000 προσομοιωμένων τιμών προσεγγίζει με μεγάλη ακρίβεια τη θεωρητική καμπύλη της τυπικής κανονικής κατανομής. Αυτό αποδεικνύει ότι το παραγόμενο δείγμα αποτελεί μια πολύ καλή εμπειρική προσέγγιση της κατανομής-στόχου.

Ερώτημα 2 (α) ii: Ανάλυση Αποδοτικότητας Αλγορίθμου

Από την εκτέλεση του κώδικα, συλλέχθηκαν τα ακόλουθα στατιστικά στοιχεία:

- Συνολικός αριθμός προτεινόμενων τιμών: `total_attempts = 13202`
- Δείγματα που έγιναν αποδεκτά μέσω του squeeze test: `squeeze_accepts = 7486`
- Δείγματα που έγιναν αποδεκτά μέσω του πλήρους ελέγχου $f(x)$: `full_accepts = 2514`
- Συνολικά αποδεκτά δείγματα: $7486 + 2514 = 10000$

Ολική πιθανότητα αποδοχής

Η εκτιμώμενη ολική πιθανότητα αποδοχής από την προσομοίωση είναι είναι ο λόγος των συνολικών αποδεκτών δειγμάτων προς τον συνολικό αριθμό των προτεινόμενων τιμών:

$$P_{\text{accepted, empirical}} = \frac{10000}{13202} \approx \boxed{0.7575}$$

Η θεωρητική πιθανότητα αποδοχής για τη rejection sampling δίνεται από τη σχέση:

$$P_{\text{accepted, theoretical}} = \frac{1}{M}, \quad \text{όπου } M = \sqrt{\frac{2e}{\pi}} \approx 1.3154 \Rightarrow \frac{1}{1.3154} \approx \boxed{0.7602}$$

Παρατηρείται ότι η εκτιμώμενη πιθανότητα αποδοχής (0.7575) είναι πολύ κοντά στη θεωρητική (0.7602), γεγονός που ενισχύει την ορθότητα της υλοποίησης.

Μέσος αριθμός προσπαθειών για μια αποδοχή

Ο αριθμός των προσπαθειών που απαιτούνται μέχρι την πρώτη επιτυχία (αποδοχή) σε μια σειρά ανεξάρτητων δοκιμών Bernoulli ακολουθεί τη Γεωμετρική κατανομή. Η πιθανότητα επιτυχίας σε κάθε δοκιμή είναι $p = P(\text{αποδοχής})$. Η μέση τιμή της Γεωμετρικής κατανομής είναι $1/p$. Επομένως, ο θεωρητικός μέσος αριθμός προσπαθειών που απαιτείται για να έχουμε μία αποδεκτή τιμή είναι:

$$\text{Μέσος Αριθμός Προσπαθειών} = \frac{1}{P(\text{αποδοχής})} = M \approx \boxed{1.3155}$$

Αυτό σημαίνει ότι, κατά μέσο όρο, χρειαζόμαστε περίπου 1.32 προσπάθειες για κάθε τιμή που τελικά θα συμπεριληφθεί στο δείγμα μας.

Πιθανότητα αποφυγής του υπολογισμού της f

Η πιθανότητα να αποφύγει κανείς τον υπολογισμό της συνάρτησης-στόχου $f(x)$ είναι η πιθανότητα ένα υποψήφιο δείγμα y να γίνει αποδεκτό από το πρώτο στάδιο, δηλαδή από το Squeeze Test, το οποίο ελέγχει εάν $u \leq \frac{s(y)}{M \cdot g(y)}$.
Έστω:

- Y είναι μια τυχαία μεταβλητή που ακολουθεί την κατανομή εισήγησης, $Y \sim g(y)$
- U είναι μια τυχαία μεταβλητή από την ομοιόμορφη κατανομή, $U \sim \mathcal{U}(0, 1)$
- A είναι το γεγονός "Αποφυγή υπολογισμού της f", το οποίο συμβαίνει όταν η συνθήκη του squeeze test ικανοποιείται. Το γεγονός A ορίζεται ως:

$$A = \left\{ U \leq \frac{s(Y)}{M \cdot g(Y)} \right\}$$

Για να βρούμε την πιθανότητα $P(A)$, θα χρησιμοποιήσουμε τον νόμο της ολικής πιθανότητας, ολοκληρώνοντας πάνω σε όλες τις πιθανές τιμές που μπορεί να πάρει η Y:

$$P(A) = \int P(A \mid Y = y) \cdot g(y)\, dy$$

Όταν έχουμε μια συγκεκριμένη τιμή $Y = y$, η συνθήκη γίνεται $u \leq \frac{s(y)}{M \cdot g(y)}$. Επειδή η u είναι ομοιόμορφα κατανεμημένη στο $[0, 1]$, η πιθανότητα $P(u \leq c)$ είναι απλώς c (για $0 \leq c \leq 1$). Άρα:

$$P(A \mid Y = y) = \frac{s(y)}{M \cdot g(y)}$$

Με αντικατάσταση στο αρχικό ολοκλήρωμα προκύπτει:

$$P(A) = \int \left[\frac{s(y)}{M \cdot g(y)}\right] \cdot g(y)\, dy = \int \frac{s(y)}{M}\, dy = \frac{1}{M} \int s(y)\, dy.$$

Αποδεικνύεται, λοιπόν, ότι η πιθανότητα να αποφύγουμε τον υπολογισμό της f είναι ίση με το εμβαδόν κάτω από την squeezing συνάρτηση $s(x)$, διαιρεμένο με τη σταθερά M.

Εν προκειμένω, είναι $M = \sqrt{2e/\pi}$ και η $s(x)$ ορίζεται από τη σχέση $s(x) \leq f(x)$. Χρησιμοποιώντας την δοθείσα ανισότητα, $e^{-x^2/2} \geq 1 - x^2/2$, και πολλαπλασιάζοντας με τον σταθεροποιητικό όρο της κανονικής κατανομής, έχουμε:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \geq \frac{1}{\sqrt{2\pi}}(1 - x^2/2)$$

Αυτή η ανισότητα ισχύει για $|x| < \sqrt{2}$, αλλιώς το δεξί μέλος γίνεται αρνητικό. Επομένως, η squeezing συνάρτηση είναι:

$$s(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \left(1 - \frac{x^2}{2}\right), & \text{αν } |x| < \sqrt{2} \\ 0, & \text{διαφορετικά} \end{cases}$$

Ακολούθως υπολογίζεται το $\int s(x)\, dx$ ως εξής:

$$\begin{aligned} \int s(x)\, dx &= \int_{-\sqrt{2}}^{+\sqrt{2}} \frac{1}{\sqrt{2\pi}} \left(1 - \frac{x^2}{2}\right) dx \\ &= \frac{1}{\sqrt{2\pi}} \left[x - \frac{x^3}{6}\right]_{-\sqrt{2}}^{+\sqrt{2}} \\ &= \frac{1}{\sqrt{2\pi}} \left[\left(\sqrt{2} - \frac{(\sqrt{2})^3}{6}\right) - \left(-\sqrt{2} + \frac{(-\sqrt{2})^3}{6}\right)\right] \\ &= \frac{4}{3\sqrt{\pi}}. \end{aligned}$$

Με αντικατάσταση στον τύπο της πιθανότητας προκύπτει:

$$P(\text{Αποφυγή } f) = \frac{1}{M} \cdot \int s(x)\, dx = \frac{1}{\sqrt{2e/\pi}} \cdot \frac{4}{3\sqrt{\pi}} = \frac{4}{3\sqrt{2e}} \approx \boxed{0.5718}.$$

Η θεωρητική πιθανότητα να αποφύγουμε τον υπολογισμό της $f(x)$ σε μία τυχαία προσπάθεια είναι περίπου 57.18%.

Με βάση τα αποτελέσματα της προσομοίωσης, ότι ο αριθμός των αποδοχών μέσω του squeeze test ήταν 7.486 επί συνόλου 13.202 προτάσεων. Άρα:

$$P(\text{Αποφυγή } f)_{\text{εμπειρική}} = \frac{7.486}{13.202} \approx \boxed{0.5670}$$

Φαίνεται ότι η εμπειρική πιθανότητα επιτυχίας που προέκυψε από την προσομοίωση είναι περίπου 0.5670, δηλαδή πολύ κοντά στην θεωρητική τιμή 0.5718, γεγονός που επιβεβαιώνει την ορθότητα της μεθόδου.

Ερώτημα 2 (α) iii: Επιθυμητά Χαρακτηριστικά Κατανομής Εισήγησης

Η επιλογή της κατανομής εισήγησης $g(x)$ είναι καθοριστική για την αποδοτικότητα της μεθόδου απόρριψης και των παραλλαγών της, όπως η squeezed rejection sampling. Ιδανικά, η $g(x)$ πρέπει να επιτρέπει εύκολη δειγματοληψία, ώστε η παραγωγή δειγμάτων να είναι υπολογιστικά απλή και ταχεία. Αν η δειγματοληψία από την $g(x)$ είναι εξίσου ή περισσότερο περίπλοκη από την $f(x)$, τότε η μέθοδος χάνει το βασικό της πλεονέκτημα.

Επιπλέον, η σταθερά M, που ικανοποιεί την ανισότητα $f(x) \leq Mg(x)$, πρέπει να είναι όσο το δυνατόν μικρότερη, καθώς η μέση πιθανότητα αποδοχής ισούται με $1/M$. Όσο μικρότερο είναι το M, τόσο λιγότερες απορρίψεις πραγματοποιούνται και τόσο ταχύτερα συλλέγονται τα δείγματα. Αυτό επιτυγχάνεται όταν το σχήμα της $Mg(x)$ προσεγγίζει καλά την $f(x)$ σε όλο το πεδίο ορισμού της. Απαραίτητο είναι επίσης η $g(x)$ να καλύπτει το πλήρες πεδίο ορισμού της $f(x)$, ώστε να είναι θετική σε όλα τα σημεία όπου και η $f(x)$ είναι θετική.

Στην περίπτωση του squeezed rejection sampling, είναι επιθυμητό να υπάρχει διαθέσιμη μια καλή συνάρτηση συμπίεσης $s(x)$, η οποία να είναι απλούστερη στον υπολογισμό από την $f(x)$ και ταυτόχρονα να την προσεγγίζει από κάτω όσο το δυνατόν καλύτερα. Όσο μεγαλύτερο τμήμα της περιοχής κάτω από την $Mg(x)$ καλύπτει η $s(x)$, τόσο περισσότερα δείγματα γίνονται αποδεκτά χωρίς τον υπολογισμό της $f(x)$.

Στην παρούσα εργασία, επιλέχθηκε ως εισήγηση η διπλή εκθετική κατανομή $g(x) = \frac{1}{2}e^{-|x|}$, καθώς παρέχει εύκολη δειγματοληψία μέσω της μεθόδου αντιστροφής, έχει συμμετρικό σχήμα που προσεγγίζει την κωδωνοειδή μορφή της κανονικής, και οδηγεί σε λογική τιμή σταθεράς $M = \sqrt{\frac{2e}{\pi}}$. Για την $s(x)$ χρησιμοποιήθηκε η προσέγγιση $s(x) = \frac{1}{\sqrt{2\pi}}\left(1 - \frac{x^2}{2}\right)$, η οποία προκύπτει από την ανισότητα $e^{-x^2/2} \geq 1 - \frac{x^2}{2}$ και είναι ορισμένη στο διάστημα $|x| < \sqrt{2}$. Η συγκεκριμένη επιλογή επέτρεψε την αποδοχή σημαντικού ποσοστού δειγμάτων μέσω του squeeze test, μειώνοντας τον συνολικό αριθμό αξιολογήσεων της $f(x)$ και βελτιώνοντας την αποδοτικότητα της μεθόδου.

Ερώτημα 2 (β) i: Εκτίμηση της Αναμενόμενης Τιμής $\mathbb{E}[\varphi(X)]$ με την Κλασική Μέθοδο Monte-Carlo

Ζητείται η εκτίμηση της μέσης τιμής $\mathbb{E}[\varphi(X)]$, όπου η συνάρτηση $\varphi(x)$ ορίζεται ως:

$$\varphi(x) = \frac{4}{1+x^2}$$

και η τυχαία μεταβλητή X ακολουθεί την ομοιόμορφη κατανομή στο διάστημα $(0,1)$, $X \sim \mathcal{U}(0,1)$. Η συνάρτηση πυκνότητας πιθανότητας (pdf) της X, $f(x)$, είναι:

$$f(x) = \begin{cases} 1, & \text{για } 0 < x < 1 \\ 0, & \text{αλλού} \end{cases}$$

Συνεπώς, η ζητούμενη μέση τιμή αντιστοιχεί στο ολοκλήρωμα:

$$\mathbb{E}[\varphi(X)] = \int_0^1 \varphi(x) f(x)\, dx = \int_0^1 \frac{4}{1+x^2} \cdot 1\, dx = \int_0^1 \frac{4}{1+x^2}\, dx$$

Η αναλυτική λύση αυτού του ολοκληρώματος είναι:

$$\int_0^1 \frac{4}{1+x^2}\, dx = 4 \cdot [\arctan(x)]_0^1 = 4 \cdot \left(\frac{\pi}{4} - 0\right) = \pi$$

Η κλασική μέθοδος Monte Carlo για την εκτίμηση ενός τέτοιου ολοκληρώματος βασίζεται στον Ισχυρό Νόμο των Μεγάλων Αριθμών [8]. Σύμφωνα με αυτόν, αν $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητα και ισόνομα κατανεμημένα δείγματα από την κατανομή της X, τότε:

$$\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) \xrightarrow{\text{σχεδόν βέβαια}} \mathbb{E}[\varphi(X)] \quad \text{καθώς } n \to \infty$$

Ο κλασικός Monte Carlo εκτιμητής για την $\mathbb{E}[\varphi(X)]$ είναι συνεπώς:

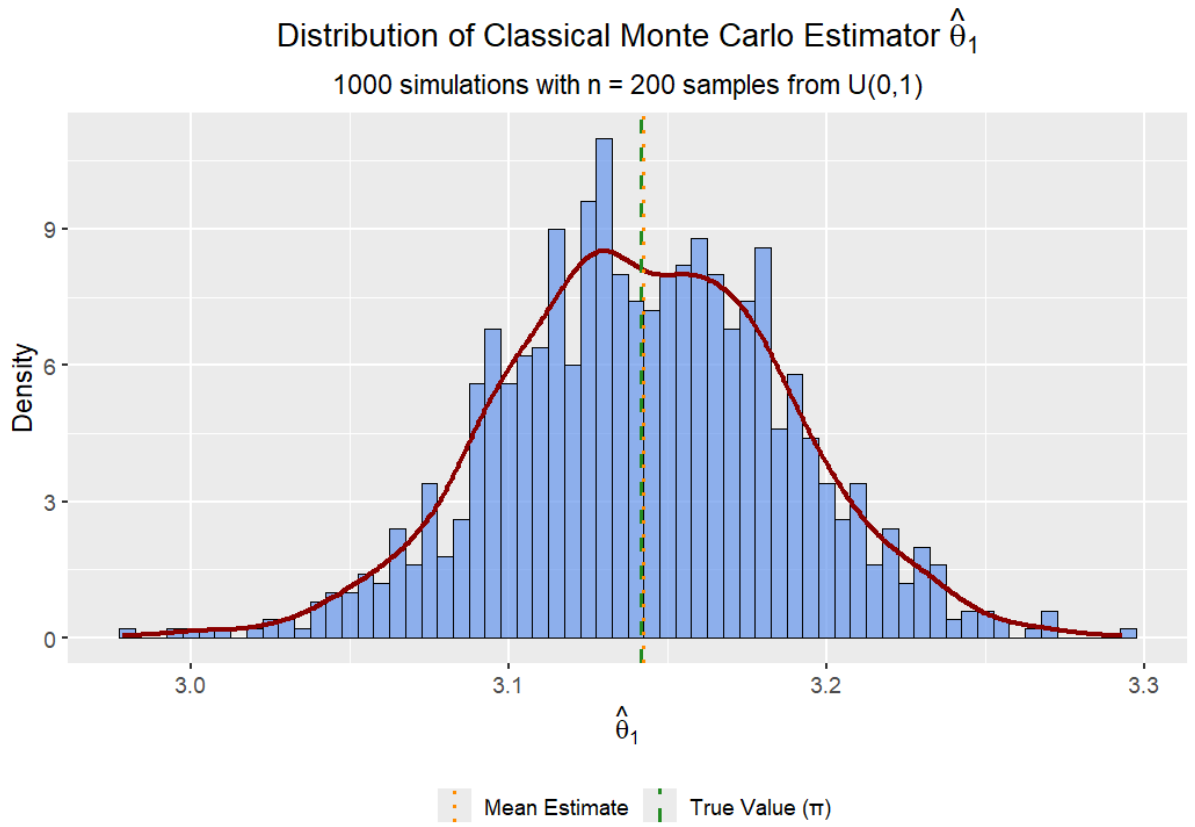
$$\widehat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i), \quad \text{όπου } X_i \sim \mathcal{U}(0,1)$$

Για να αξιολογήσουμε την ακρίβεια του εκτιμητή $\widehat{\theta}_1$, μπορούμε να επαναλάβουμε την προσομοίωση της διαδικασίας εκτίμησης πολλές φορές και να υπολογίσουμε την τυπική απόκλιση των παραγόμενων εκτιμήσεων $\widehat{\theta}_1$. Αυτή η τυπική απόκλιση αποτελεί εκτίμηση του τυπικού σφάλματος (standard error) του εκτιμητή.

Η εκτίμηση της E[φ(X)] και του τυπικού σφάλματος του εκτιμητή $\widehat{\theta}_1$ υλοποιείται με τον ακόλουθο κώδικα, όπου ορίζεται η συνάρτηση `phi(x)` και χρησιμοποιείται η συνάρτηση `replicate` για την εκτέλεση 1000 προσομοιώσεων. Σε κάθε προσομοίωση παράγονται $n = 200$ δείγματα `x_samples` από την `runif(n)` και υπολογίζεται ο μέσος όρος `mean(phi(x_samples))`, ο οποίος αποτελεί μία τιμή του $\hat{\theta}_1$. Στη συνέχεια υπολογίζεται η `sd(theta_hat_1_values)` για την εκτίμηση του τυπικού σφάλματος. Τέλος κατασκευάζεται το ιστόγραμμα και η καμπύλη πυκνότητας των τιμών $\hat{\theta}_1$, στο οποίο προστίθενται η πραγματική τιμή του π και ο μέσος όρος των εκτιμήσεων.

```r
# Classical Monte Carlo Estimator for Eφ[(X)]
# Define the function φ(x)
phi <- function(x) {
  return(4 / (1 + x^2))
}

```

```r
# Number of simulations (Monte Carlo estimates)
num_simulations <- 1000

# Sample size per estimation
n <- 200

# Use replicate to generate 1000 estimates of θ₁
set.seed(123)
theta_hat_1_values <- replicate(num_simulations, {
  x <- runif(n)
  mean(phi(x))
})

# Estimate the standard error: standard deviation of the estimates
se_theta_hat_1 <- sd(theta_hat_1_values)

# True value of the integral (Eφ[(X)]) = π
true_value <- pi
mean_estimate <- mean(theta_hat_1_values)

# Create histogram + density + vertical lines

df_theta1 <- data.frame(theta_hat_1 = theta_hat_1_values)

ggplot(df_theta1, aes(x = theta_hat_1)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.005, fill = "cornflowerblue",
    color = "black", alpha = 0.7) +
  geom_density(color = "darkred", size = 1.2, alpha = 0.9) +
  geom_vline(aes(xintercept = true_value, color = "True Value π()"), linetype = "
    dashed", size = 1) +
  geom_vline(aes(xintercept = mean_estimate, color = "Mean Estimate"), linetype = "
    dotted", size = 1) +
  scale_color_manual(name = "", values = c("True Value π()" = "forestgreen", "Mean
    Estimate" = "darkorange")) +
  labs(
    title = expression(paste("Distribution of Classical Monte Carlo Estimator ",
    hat(theta)[1])),
    subtitle = paste0(num_simulations, " simulations with n = ", n, " samples from
    U(0,1)"),
    x = expression(hat(theta)[1]),
    y = "Density" ) +
  theme_gray(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "bottom")

# Print numerical results
cat("Estimated standard error of θ₁:", round(se_theta_hat_1, 5), "\n")
cat("Mean of simulated θ₁ values:", round(mean_estimate, 5), "\n")
cat("True value π():", round(true_value, 5), "\n")
```



Σχήμα 7. *Κατανομή του κλασικού εκτιμητή Monte Carlo $\hat{\theta}_1$ για την εκτίμηση του π, με βάση 1000 επαναλήψεις δειγματοληψίας μεγέθους $n = 200$ από την $U(0, 1)$. Με πράσινη διακεκομμένη γραμμή σημειώνεται η πραγματική τιμή π και με πορτοκαλί η μέση εκτίμηση του εκτιμητή.*

Από την εκτέλεση του παραπάνω κώδικα, λαμβάνεται το γράφημα που παρουσιάζεται στο Σχήμα 7 καθώς και τα ακόλουθα αριθμητικά αποτελέσματα:

- Εκτιμώμενο τυπικό σφάλμα του $\widehat{\theta}_1$: 0.04524
- Μέσος όρος των προσομοιωμένων τιμών $\widehat{\theta}_1$: 3.14247
- Πραγματική τιμή (π): 3.14159

Το εκτιμώμενο τυπικό σφάλμα του $\widehat{\theta}_1$ είναι μικρό, ίσο με 0.04524. Αυτή η τιμή ποσοτικοποιεί την αβεβαιότητα που σχετίζεται με τον εκτιμητή όταν χρησιμοποιείται μέγεθος δείγματος n=200. Μικρότερο τυπικό σφάλμα υποδηλώνει μεγαλύτερη ακρίβεια του εκτιμητή.

Επίσης, ο μέσος όρος των 1000 προσομοιωμένων τιμών του εκτιμητή $\widehat{\theta}_1$ (3.14247) είναι πολύ κοντά στην αληθινή τιμή του ολοκληρώματος, $\pi \approx 3.14159$. Αυτό επιβεβαιώνει την αμεροληψία του κλασικού Monte-Carlo εκτιμητή για επαρκώς μεγάλο αριθμό δειγμάτων ανά εκτίμηση (n) και μεγάλο αριθμό προσομοιώσεων. Το ιστόγραμμα και η καμπύλη πυκνότητας δείχνουν ότι οι τιμές του $\widehat{\theta}_1$ κατανέμονται συμμετρικά γύρω από την πραγματική τιμή, όπως αναμένεται για έναν μέσο όρο από ανεξάρτητα δείγματα (λόγω του Κεντρικού Οριακού Θεωρήματος, η κατανομή του $\widehat{\theta}_1$ προσεγγίζει την κανονική).

Ερώτημα 2 (β) ii: Εκτίμηση της $\mathbb{E}[\varphi(X)]$ με Δειγματοληψία Σπουδαιότητας

Η μέθοδος σπουδαιότητας (importance sampling) αποτελεί μία τεχνική μείωσης της διακύμανσης, η οποία μπορεί να βελτιώσει την αποδοτικότητα της προσομοίωσης Monte Carlo. Το αρχικό ολοκλήρωμα $\theta = \mathbb{E}[\phi(X)] = \int \phi(x)f(x)dx$ μπορεί να γραφτεί ως εξής, πολλαπλασιάζοντας και διαιρώντας με μια νέα συνάρτηση πυκνότητας πιθανότητας $g(x)$, την κατανομή σπουδαιότητας (importance distribution) [6]:

$$\theta = \int \left[\frac{\phi(x)f(x)}{g(x)}\right] g(x)dx = \mathbb{E}_g[\psi(X)],$$

όπου $\psi(x) = \frac{\phi(x)f(x)}{g(x)}$ είναι το βάρος σπουδαιότητας (importance weight). Η εκτίμηση του θ γίνεται τώρα μέσω του εκτιμητή:

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} \psi(X_i),$$

όπου τα δείγματα $X_1, \ldots, X_n$ παράγονται πλέον από την κατανομή $g(x)$ και όχι από την αρχική $f(x)$.

Η διασπορά του νέου εκτιμητή είναι $V_g[\hat{\theta}_2] = \frac{1}{n}V_g[\psi(X)]$. Η αποτελεσματικότητα της μεθόδου έγκειται στην επιλογή μιας κατάλληλης $g(x)$ που να ελαχιστοποιεί τη διασπορά $V_g[\psi(X)]$. Θεωρητικά, η βέλτιστη $g(x)$ είναι ανάλογη του $\phi(x)f(x)$. Εν προκειμένω, $f(x) = 1$, οπότε αναζητούμε μια $g(x)$ που να "μιμείται" τη συμπεριφορά της $\phi(x)$ στο διάστημα $[0, 1]$.

Δίνεται η $g(x) = \frac{1}{3}(4 - 2x)$, η οποία είναι μια γραμμική συνάρτηση που προσπαθεί να προσεγγίσει την καμπύλη της $\phi(x)$. Η $\psi(x)$ υπολογίζεται ως:

$$\psi(x) = \frac{\phi(x)}{g(x)} = \frac{4}{1+x^2} \cdot \frac{3}{4-2x} = \frac{12}{(1+x^2)(4-2x)}.$$

Αν η $g(x)$ αποτελεί καλή προσέγγιση της $\phi(x)$, τότε ο λόγος τους, δηλαδή η $\psi(x)$, θα είναι σχεδόν σταθερός, με αποτέλεσμα ο εκτιμητής να έχει πολύ μικρή διασπορά.

Ακολούθως παρατίθεται ο κώδικας για την υλοποίηση του εκτιμητή με δειγματοληψία σπουδαιότητας. Αρχικά ορίζονται οι συναρτήσεις `phi(x)` και `g_pdf(x)`. Δεδομένου ότι δεν υπάρχει ενσωματωμένη συνάρτηση για την παραγωγή τιμών από την g(x), χρησιμοποιήθηκε η Μέθοδος του Αντίστροφου Μετασχηματισμού (Inversion Method). Υπολογίστηκε η αθροιστική συνάρτηση κατανομής:

$$G(x) = \int_0^x g(t)\,dt = \frac{1}{3}(4x - x^2),$$

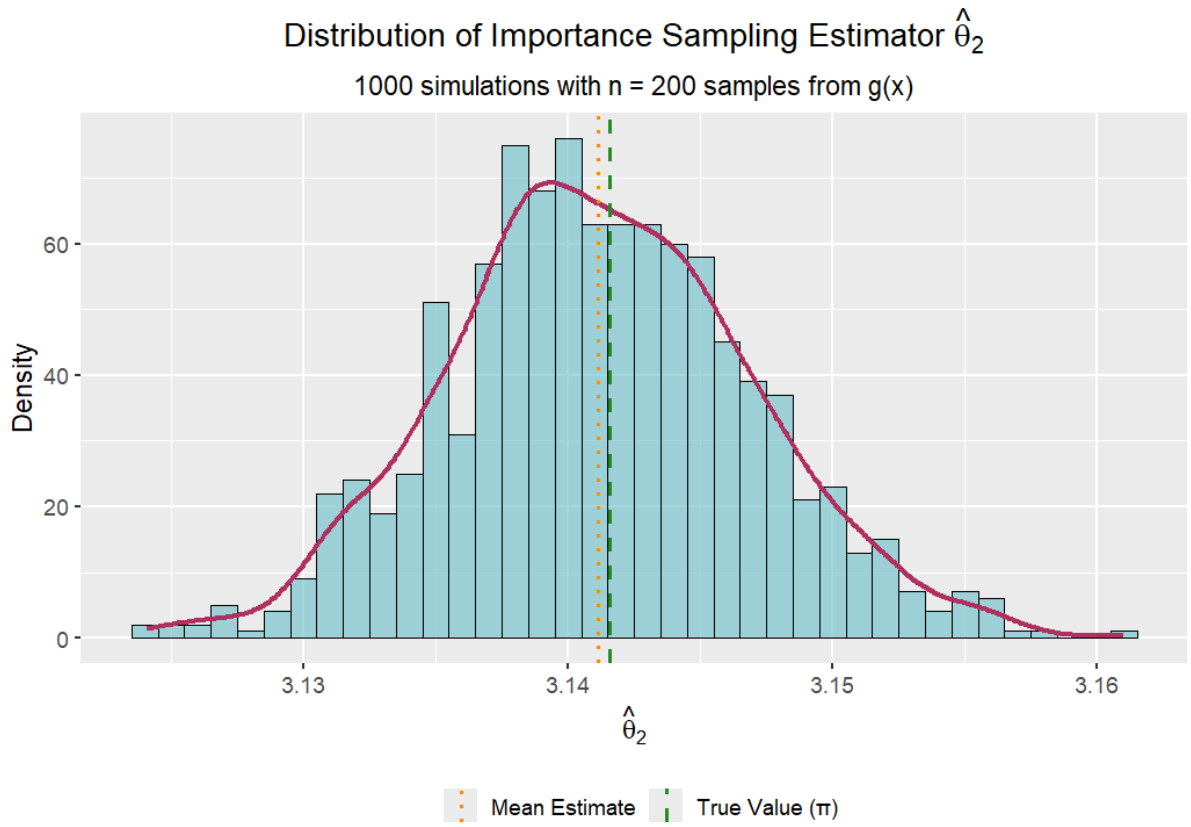
λύθηκε η εξίσωση $G(x) = u$ ως προς x, και προέκυψε ο τύπος παραγωγής δειγμάτων:

$$x = G^{-1}(u) = 2 - \sqrt{4 - 3u}, \quad \text{όπου } u \sim U(0, 1).$$

Αυτό υλοποιήθηκε στη συνάρτηση `generate_from_g`. Ορίστηκε η συνάρτηση `psi(x)` που υπολογίζει τα βάρη σπουδαιότητας. Για την προσομοίωση χρησιμοποιήθηκε η `replicate` για την παραγωγή 1000 τιμών του $\hat{\theta}_2$. Σε κάθε επανάληψη, παράγεται ένα δείγμα $n = 200$ από την $g(x)$, υπολογίζονται τα αντίστοιχα βάρη $\psi(x_i)$, και εξάγεται ο μέσος όρος τους, ο οποίος αποτελεί μια τιμή του $\hat{\theta}_2$. Τέλος, υπολογίστηκε το τυπικό σφάλμα του $\hat{\theta}_2$ και σχεδιάστηκε το ιστόγραμμα και η καμπύλη πυκνότητας των τιμών $\hat{\theta}_2$, μαζί με την πραγματική τιμή και τον μέσο όρο των εκτιμήσεων.

```r
# Importance Sampling Estimator for Eφ[(X)]
# Define phi(x)
phi <- function(x) 4 / (1 + x^2)

# Importance sampling density g(x)
```

```r
g_pdf <- function(x) (1/3)*(4 - 2*x)

# Inverse CDF sampling from g(x)
generate_from_g <- function(n) {
  u <- runif(n)
  2 - sqrt(4 - 3 * u)}

# Importance sampling weight ψ(x)
psi <- function(x) phi(x) / g_pdf(x)

# Simulation parameters
num_sim <- 1000
n <- 200

# Generate estimates of θ₂
set.seed(456)
theta_hat_2 <- replicate(num_sim, {
  x <- generate_from_g(n)
  mean(psi(x))})

# Compute standard error and mean estimate
se_theta_hat_2 <- sd(theta_hat_2)
mean_estimate <- mean(theta_hat_2)
true_value <- pi

# Plot distribution of θ₂
df <- data.frame(theta = theta_hat_2)

ggplot(df, aes(x = theta)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.001, fill = "cadetblue3", color
     = "black", alpha = 0.7) +
  geom_density(color = "maroon", size = 1.2) +
  geom_vline(xintercept = true_value, color = "forestgreen", linetype = "dashed",
    size = 1) +
  geom_vline(xintercept = mean_estimate, color = "darkorange", linetype = "dotted",
     size = 1) +
  labs(
    title = expression(paste("Distribution of Importance Sampling Estimator ", hat(
    theta)[2])),
    subtitle = paste(num_sim, "simulations with n =", n, "samples from g(x)"),
    x = expression(hat(theta)[2]), y = "Density"
  ) +
  theme_gray(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = "none")

# Print results
cat("Standard error of θ₂:", round(se_theta_hat_2, 5), "\n")
cat("Mean estimate:", round(mean_estimate, 5), "\n")
cat("True value π():", round(true_value, 5), "\n")
```



Σχήμα 8. *Κατανομή του εκτιμητή δειγματοληψίας σπουδαιότητας $\hat{\theta}_2$ για την εκτίμηση του π, με βάση 1000 επαναλήψεις δειγματοληψίας μεγέθους $n = 200$ από την κατανομή σπουδαιότητας $g(x) = \frac{1}{3}(4 - 2x)$. Με πράσινη διακεκομμένη γραμμή σημειώνεται η αληθινή τιμή π και με πορτοκαλί η μέση εκτίμηση.*

Από την εκτέλεση του παραπάνω κώδικα, λαμβάνεται το γράφημα που παρουσιάζεται στο Σχήμα 8 καθώς και τα ακόλουθα αριθμητικά αποτελέσματα:

- Εκτιμώμενο τυπικό σφάλμα του $\widehat{\theta}_2$: 0.00575
- Μέσος όρος των προσομοιωμένων τιμών $\widehat{\theta}_2$: 3.14114

Το εκτιμώμενο τυπικό σφάλμα του $\widehat{\theta}_2$ είναι σχεδόν μηδενικό, ίσο με 0.00575. Επίσης, ο μέσος όρος των 1000 προσομοιωμένων τιμών του εκτιμητή $\widehat{\theta}_2$ (3.14114) είναι πολύ κοντά στην αληθινή τιμή του π, υποδεικνύοντας ότι και αυτός ο εκτιμητής είναι αμερόληπτος (όπως αναμένεται θεωρητικά).

Το τυπικό σφάλμα του Κλασικού Monte Carlo εκτιμητή $\widehat{\theta}_1$ είναι $\text{SE}_{\text{MC}} = 0.04481$, ενώ του εκτιμητή με Δειγματοληψία Σπουδαιότητας $\widehat{\theta}_2$ είναι $\text{SE}_{\text{IS}} = 0.00575$. Είναι εμφανές ότι το τυπικό σφάλμα του εκτιμητή με δειγματοληψία σπουδαιότητας $\widehat{\theta}_2$ είναι σημαντικά μικρότερο από αυτό του κλασικού Monte Carlo εκτιμητή $\widehat{\theta}_1$. Συγκεκριμένα, η μείωση του τυπικού σφάλματος ανέρχεται περίπου σε $\frac{0.04481 - 0.00575}{0.04481} \cdot 100\% \approx 87.18\%$

Η δειγματοληψία σπουδαιότητας συνέβαλε αισθητά στη μείωση της διακύμανσης του εκτιμητή. Αυτό σημαίνει ότι για τον ίδιο αριθμό δειγμάτων n, ο εκτιμητής $\widehat{\theta}_2$ παρέχει, κατά μέσο όρο, πιο ακριβείς εκτιμήσεις της ζητούμενης αναμενόμενης τιμής $\mathbb{E}[\varphi(X)]$. Η επιτυχία αυτή οφείλεται πιθανότατα στο γεγονός ότι η $g(x)$ κατανέμει τα δείγματά της με τρόπο που ευνοεί τις περιοχές του διαστήματος $(0, 1)$ όπου η συνάρτηση $\varphi(x)f(x)$ (στην περίπτωσή μας απλά $\varphi(x)$) έχει μεγαλύτερη συνεισφορά στο ολοκλήρωμα. Η γραμμική μορφή της $g(x) = \frac{1}{3}(4 - 2x)$ δίνει μεγαλύτερη πιθανότητα σε μικρότερες τιμές του x, όπου η $\varphi(x) = \frac{4}{1+x^2}$ παίρνει μεγαλύτερες τιμές, σε αντίθεση με την ομοιόμορφη $f(x)$, η οποία αποδίδει ίση πιθανότητα σε όλο το διάστημα.

Άσκηση 3

Εκτίμηση Παραμέτρου Μίξης Εκθετικών Κατανομών με τον Αλγόριθμο ΕΜ

Εξετάζεται ένα μοντέλο στο οποίο οι παρατηρήσεις X προέρχονται από μία μίξη δύο εκθετικών κατανομών. Η επιλογή της κατανομής καθορίζεται από μία λανθάνουσα (μη παρατηρούμενη) τυχαία μεταβλητή Z, η οποία ακολουθεί την κατανομή Bernoulli με άγνωστη παράμετρο p, $Z \sim \text{Bernoulli}(p)$.

- Αν $Z = 1$ (με πιθανότητα p), τότε $X \mid Z = 1 \sim \text{Exp}(\lambda_1), \quad \lambda_1 = 1$
- Αν $Z = 0$ (με πιθανότητα $1 - p$), τότε $X \mid Z = 0 \sim \text{Exp}(\lambda_0), \quad \lambda_0 = 5$

Δεδομένου ενός συνόλου ανεξάρτητων και ισόνομα κατανεμημένων παρατηρήσεων $X = (x_1, x_2, \ldots, x_n)$, όπου οι αντίστοιχες τιμές των λανθανουσών μεταβλητών $Z = (z_1, z_2, \ldots, z_n)$ είναι άγνωστες, ο στόχος είναι η εκτίμηση της άγνωστης παραμέτρου p. Για τον σκοπό αυτό χρησιμοποιείται ο αλγόριθμος ΕΜ, ο οποίος αποτελείται από δύο βήματα που επαναλαμβάνονται μέχρι τη σύγκλιση. Στο βήμα-Ε (Expectation step) υπολογίζεται η αναμενόμενη τιμή των λανθανουσών μεταβλητών Z_i, δεδομένων των παρατηρούμενων X_i και της τρέχουσας εκτίμησης του p. Στο βήμα-Μ (Maximization step) μεγιστοποιείται η αναμενόμενη πλήρης λογαριθμική πιθανοφάνεια ως προς την παράμετρο p [9].

Το πλήρες σύνολο δεδομένων είναι $(X, Z) = \{(X_1, Z_1), \ldots, (X_n, Z_n)\}$. Η συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) για μια μεμονωμένη παρατήρηση (X_i, Z_i) είναι:

$$f(x_i, z_i \mid p) = f(x_i \mid z_i) \cdot P(z_i \mid p).$$

Η παραπάνω σχέση μπορεί να γραφεί σε ενιαία μορφή χρησιμοποιώντας την ιδιότητα της Bernoulli (όπου $z_i \in \{0, 1\}$):

$$f(x_i, z_i \mid p) = [p \cdot f(x_i \mid Z_i = 1)]^{z_i} \cdot [(1 - p) \cdot f(x_i \mid Z_i = 0)]^{1 - z_i}$$

Αντικαθιστώντας τις σ.π.π. των εκθετικών κατανομών $f(x; \lambda) = \lambda e^{-\lambda x}$:

$$f(x_i, z_i \mid p) = \left[p \cdot 1 \cdot e^{-x_i}\right]^{z_i} \cdot \left[(1 - p) \cdot 5 \cdot e^{-5x_i}\right]^{1 - z_i}$$

Η πλήρης πιθανοφάνεια για όλο το δείγμα είναι το γινόμενο των επιμέρους όρων:

$$L(p; X, Z) = \prod_{i=1}^{n} f(x_i, z_i \mid p)$$

Η λογαριθμική πλήρης πιθανοφάνεια είναι:

$$\ell(p; X, Z) = \log L(p; X, Z) = \sum_{i=1}^{n} \left[z_i \cdot \log(p \cdot e^{-x_i}) + (1 - z_i) \cdot \log((1 - p) \cdot 5e^{-5x_i})\right] \Longrightarrow$$

$$\ell(p; X, Z) = \sum_{i=1}^{n} \left[z_i \cdot (\log p - x_i) + (1 - z_i) \cdot (\log(1 - p) + \log 5 - 5x_i)\right]$$

Ομαδοποιώντας τους όρους που εξαρτώνται από την παράμετρο p:

$$\ell(p; X, Z) = \log p \cdot \sum_{i=1}^{n} z_i + \log(1 - p) \cdot \sum_{i=1}^{n} (1 - z_i) + C,$$

όπου C είναι σταθεροί όροι (δεν εξαρτώνται από το p).

Βήμα Ε (Expectation Step)

Στο βήμα-Ε της r-οστής επανάληψης, υπολογίζεται η αναμενόμενη τιμή της πλήρους λογαριθμικής πιθανοφάνειας. Η προσδοκία υπολογίζεται ως προς τη δεσμευμένη κατανομή των λανθανουσών μεταβλητών Z δεδομένων των παρατηρήσεων X και της τρέχουσας εκτίμησης της παραμέτρου $p^{(r)}$. Ορίζεται η συνάρτηση:

$$Q(p, p^{(r)}) = \mathbb{E}_{Z|X,p^{(r)}} \left[\ell(p; X, Z)\right].$$

Λόγω της γραμμικότητας της αναμενόμενης τιμής, χρειαζόμαστε μόνο τις αναμενόμενες τιμές των z_i:

$$\mathbb{E}[z_i \mid X, p^{(r)}] = \mathbb{E}[z_i \mid x_i, p^{(r)}],$$

αφού τα (x_i, z_i) είναι ανεξάρτητα μεταξύ τους. Ορίζουμε:

$$w_i^{(r)} = \mathbb{E}[z_i \mid x_i, p^{(r)}],$$

δηλαδή τη δεσμευμένη πιθανότητα η i-οστή παρατήρηση να προέρχεται από την πρώτη συνιστώσα ($Z_i = 1$), δεδομένου του x_i και της τρέχουσας εκτίμησης $p^{(r)}$. Με χρήση του κανόνα του Bayes:

$$w_i^{(r)} = P(Z_i = 1 \mid x_i, p^{(r)}) = \frac{f(x_i \mid Z_i = 1) \cdot P(Z_i = 1 \mid p^{(r)})}{f(x_i \mid p^{(r)})}$$

Ο παρονομαστής $f(x_i \mid p^{(r)})$ είναι η περιθώρια πυκνότητα πιθανότητας, που προκύπτει από τον νόμο της ολικής πιθανότητας:

$$f(x_i \mid p^{(r)}) = f(x_i \mid Z_i = 1) \cdot P(Z_i = 1 \mid p^{(r)}) + f(x_i \mid Z_i = 0) \cdot P(Z_i = 0 \mid p^{(r)}) \Longrightarrow$$

$$f(x_i \mid p^{(r)}) = \left(1 \cdot e^{-x_i}\right) \cdot p^{(r)} + \left(5 \cdot e^{-5x_i}\right) \cdot (1 - p^{(r)})$$

Επομένως, το βάρος $w_i^{(r)}$ δίνεται από τη σχέση:

$$w_i^{(r)} = \frac{p^{(r)} \cdot e^{-x_i}}{p^{(r)} \cdot e^{-x_i} + (1 - p^{(r)}) \cdot 5e^{-5x_i}}$$

Τελικά, η συνάρτηση $Q(p, p^{(r)})$ γράφεται ως εξής:

$$Q(p, p^{(r)}) = \mathbb{E}\left[\log p \cdot \sum z_i + \log(1 - p) \cdot \sum (1 - z_i) + C\right]$$

$$Q(p, p^{(r)}) = \log p \cdot \sum w_i^{(r)} + \log(1 - p) \cdot \sum (1 - w_i^{(r)}) + \text{const}$$

Βήμα Μ (Maximization Step)

Στο βήμα-Μ, μεγιστοποιούμε τη συνάρτηση $Q(p, p^{(r)})$ ως προς p ώστε να υπολογίσουμε την επόμενη εκτίμηση $p^{(r+1)}$:

$$p^{(r+1)} = \arg\max_p Q(p, p^{(r)})$$

Παραγωγίζουμε ως προς p και εξισώνουμε με μηδέν:

$$\frac{\partial Q}{\partial p} = \frac{1}{p} \sum_{i=1}^{n} w_i^{(r)} - \frac{1}{1 - p} \sum_{i=1}^{n} (1 - w_i^{(r)}) = 0 \Longrightarrow$$

$$\frac{1}{p}\sum_{i=1}^{n} w_i^{(r)} = \frac{1}{1-p}\left(n - \sum_{i=1}^{n} w_i^{(r)}\right) \Longrightarrow$$

$$(1-p)\sum_{i=1}^{n} w_i^{(r)} = p\left(n - \sum_{i=1}^{n} w_i^{(r)}\right) \Longrightarrow$$

$$\sum_{i=1}^{n} w_i^{(r)} = pn.$$

Άρα η νέα εκτίμηση $p^{(r+1)}$ προκύπτει από τον ακόλουθο κανόνα ενημέρωσης:

$$p^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} w_i^{(r)}$$

Το αποτέλεσμα αυτό ερμηνεύεται ως εξής: η νέα εκτίμηση της πιθανότητας p είναι απλώς ο μέσος όρος των δεσμευμένων πιθανοτήτων (ή "βαρών") κάθε παρατήρηση να προέρχεται από την πρώτη συνιστώσα.

Υλοποίηση σε R

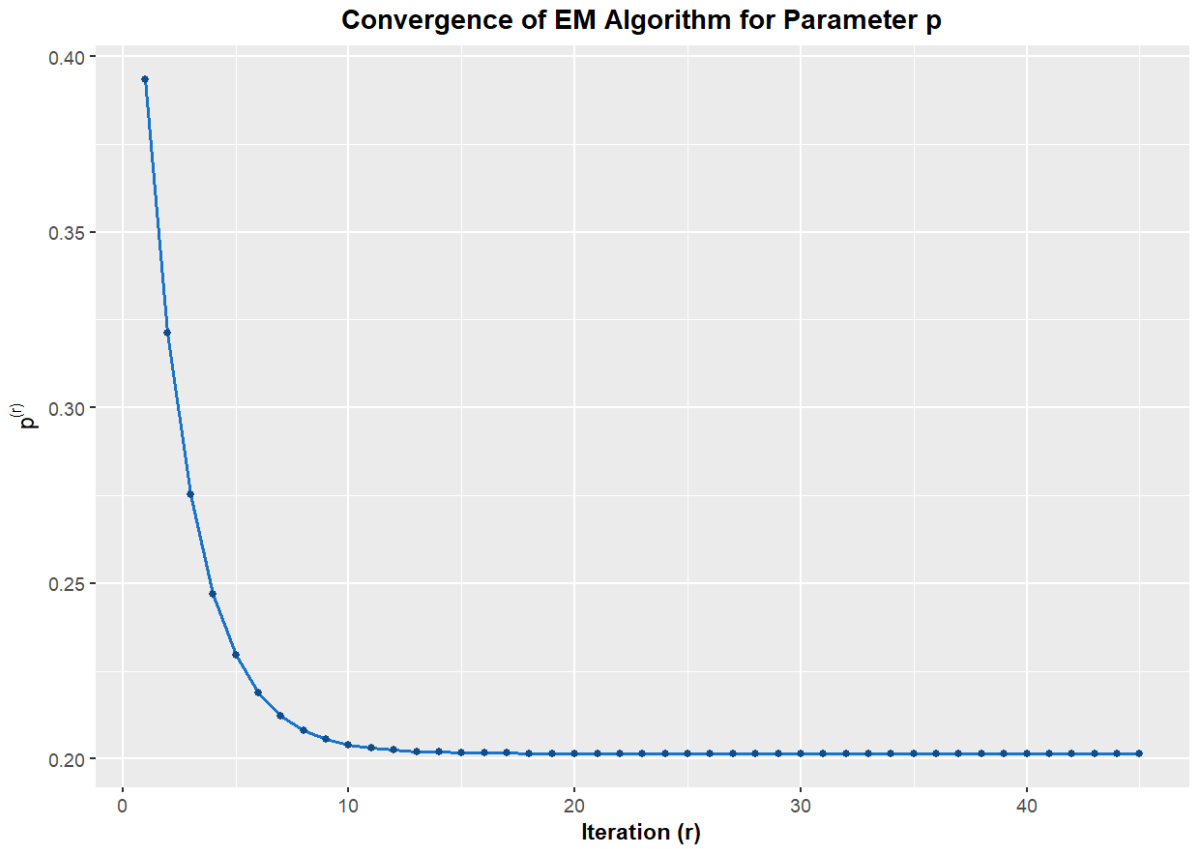
Ακολουθεί η υλοποίηση του αλγορίθμου ΕΜ στο πρόβλημα μίξης δύο εκθετικών κατανομών σύμφωνα με τη λογική που περιγράφηκε παραπάνω. Ο αλγόριθμος λειτουργεί ως εξής:

1. **Αρχικοποίηση**: Δίνεται μια αρχική τιμή στην παράμετρο, $p^{(0)}$.
2. **E-Step**: Με βάση την τρέχουσα εκτίμηση $p^{(r)}$, υπολογίζεται για κάθε παρατήρηση x_i η δεσμευμένη πιθανότητα να προέρχεται από την πρώτη εκθετική κατανομή. Αυτές οι πιθανότητες, δηλαδή τα "βάρη" $w_i^{(r)}$, εκφράζουν την αναμενόμενη τιμή της λανθάνουσας μεταβλητής Z_i.
3. **M-Step**: Η παράμετρος p ενημερώνεται μεγιστοποιώντας την αναμενόμενη τιμή της λογαριθμικής πιθανοφάνειας των πλήρων δεδομένων. Αυτό ισοδυναμεί με τον υπολογισμό του μέσου όρου των βαρών w_i που βρέθηκαν στο προηγούμενο βήμα, οδηγώντας στη νέα εκτίμηση $p^{(r+1)}$.

Τα βήματα `E` και `M` επαναλαμβάνονται μέχρι η διαφορά μεταξύ δύο διαδοχικών εκτιμήσεων της παραμέτρου `p` να γίνει αμελητέα ($\left|p^{(r+1)} - p^{(r)}\right| \leq 10^{-10}$), οπότε και θεωρείται ότι ο αλγόριθμος έχει συγκλίνει.

```r
# Load the data
x <- readRDS("data3em.rds")

# EM algorithm for estimating parameter p
em_algorithm <- function(x, p_initial = 0.5, tol = 1e-10, max_iter = 1000) {

  # Initialize parameter p with the value p(0)
  p_current <- p_initial

  # Precompute densities for efficiency:
  # f(x_i | Z_i = 1) = 1 * exp(-x_i)
  # f(x_i | Z_i = 0) = 5 * exp(-5x_i)
  f_z1 <- dexp(x, rate = 1)
  f_z0 <- dexp(x, rate = 5)

  # Vector to store the history of p estimates
  p_history <- numeric(max_iter)
```

```r

  # Begin EM iterations
  for (r in 1:max_iter) {

    # E-Step: Compute the weights wi(r) = P(Zi=1 | xi,p(r))
    numerator <- p_current * f_z1
    denominator <- numerator + (1 - p_current) * f_z0
    w <- numerator / denominator

    # M-Step: Update parameter p: p(r+1) = (1/n) * sum wi(r)
    p_new <- mean(w)

    # Store new value and check for convergence
    p_history[r] <- p_new
    if (abs(p_new - p_current) <= tol) {
      p_history <- p_history[1:r] # Trim the history vector
      break
    }

    # Update for the next iteration
    p_current <- p_new
  }

  return(list(p_hat = p_new, iterations = r, p_history = p_history))
}

# Run the EM algorithm
result <- em_algorithm(x, p_initial = 0.5)

# Print final results
cat("Estimated p (p_hat):", format(result$p_hat, digits = 11), "\n")
cat("EM iterations:", result$iterations, "\n")

# Visualization of convergence

df_plot <- data.frame(iteration = 1:result$iterations, p_value = result$p_history)

ggplot(df_plot, aes(x = iteration, y = p_value)) +
  geom_line(color = "dodgerblue3", linewidth = 1) +
  geom_point(color = "dodgerblue4", size = 2) +
  labs(title = "Convergence of EM Algorithm for Parameter p",
       x = "Iteration (r)",
       y = expression(p^"(r)")) +
  theme_gray(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold")
  )
```



Σχήμα 9. *Σύγκλιση του αλγορίθμου ΕΜ για την εκτίμηση της παραμέτρου* $\hat{p}^{(r)}$.

Ο αλγόριθμος ΕΜ εφαρμόστηκε στα δεδομένα του αρχείου `data3em.rds` με αρχική εκτίμηση $p^{(0)} = 0.5$ και λήφθηκε το γράφημα του Σχήματος 9, καθώς και τα ακόλουθα αποτελέσματα:

- Εκτιμώμενη τιμή της παραμέτρου: $\hat{p} = 0.201536$
- Αριθμός επαναλήψεων μέχρι τη σύγκλιση: 45

Στο Σχήμα 9 φαίνεται ότι ο αλγόριθμος ΕΜ συγκλίνει γρήγορα και η εκτίμηση της παραμέτρου p μετατοπίζεται σημαντικά προς την τελική τιμή από τις πρώτες κιόλας επαναλήψεις. Η τιμή $\hat{p} = 0.201536$ αποτελεί την εκτίμηση μέγιστης πιθανοφάνειας (MLE) της άγνωστης παραμέτρου μίξης p, δεδομένων των παρατηρήσεων και του υποκείμενου μοντέλου. Η εκτίμηση υποδηλώνει ότι περίπου το 20.15% των παρατηρήσεων προέρχεται από την εκθετική κατανομή με παράμετρο $\lambda_1 = 1$, ενώ το υπόλοιπο 79.85% από την κατανομή με $\lambda_0 = 5$. Η σχετικά γρήγορη σύγκλιση (μόλις 45 επαναλήψεις) για πολύ μικρό `tol` επιβεβαιώνει την αποτελεσματικότητα του αλγορίθμου ΕΜ για το συγκεκριμένο πρόβλημα μίξης.

Άσκηση 4

Ερώτημα 4 (α) - Επιλογή Υπομοντέλων με Κριτήριο AIC

Δίνεται ένα πρόβλημα επιλογής μεταβλητών (Variable Selection) στα πλαίσια της πολλαπλής γραμμικής παλινδρόμησης (Multiple Linear Regression). Στόχος είναι η μοντελοποίηση της μεταβλητής απόκρισης `medv` (μέση αξία ιδιοκατοίκησης) χρησιμοποιώντας έναν βέλτιστο συνδυασμό από τις $p = 13$ διαθέσιμες επεξηγηματικές μεταβλητές των δεδομένων `Boston`.

Το γενικό μοντέλο παλινδρόμησης έχει τη μορφή:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

όπου είναι η `medv` και X_j είναι οι επεξηγηματικές μεταβλητές. Το ζητούμενο είναι να επιλέξουμε ποιες από αυτές τις μεταβλητές θα συμπεριληφθούν στο τελικό μοντέλο. Κάθε υποσύνολο των p μεταβλητών ορίζει ένα διαφορετικό μοντέλο. Ο συνολικός αριθμός των πιθανών μοντέλων που μπορούν να κατασκευαστούν είναι $2^p = 2^13 = 8.192$.

Για την επιλογή του "καλύτερου" μοντέλου, εφαρμόζουμε την αρχή της φειδωλότητας (Parsimony), η οποία επιδιώκει την ισορροπία μεταξύ της καλής προσαρμογής (Goodness of Fit) και της πολυπλοκότητας του μοντέλου [10]. Ένα μοντέλο με περισσότερες μεταβλητές θα έχει πάντα καλύτερη προσαρμογή στα δεδομένα εκπαίδευσης (μικρότερο Άθροισμα Τετραγώνων Υπολοίπων - RSS), αλλά κινδυνεύει από υπερπροσαρμογή (overfitting) και χαμηλή γενικευσιμότητα.

Το κριτήριο που χρησιμοποιούμε για την αξιολόγηση αυτής της ισορροπίας είναι το Κριτήριο Πληροφορίας του Akaike (Akaike Information Criterion, AIC). Το AIC ορίζεται ως:

$$\text{AIC} = -2 \cdot \ln(\mathcal{L}) + 2k$$

όπου $\mathcal{L}$ είναι η μέγιστη τιμή της συνάρτησης πιθανοφάνειας του μοντέλου και k είναι ο αριθμός των παραμέτρων που εκτιμήθηκαν στο μοντέλο (συμπεριλαμβανομένης της σταθεράς, $k = s + 1$, όπου s ο αριθμός των μεταβλητών).

Για γραμμικά μοντέλα με κανονικά κατανεμημένα σφάλματα, το AIC μπορεί να εκφραστεί (παραλείποντας σταθερούς όρους) ως:

$$\text{AIC} = n \cdot \ln\left(\frac{\text{RSS}}{n}\right) + 2k'$$

όπου n είναι ο αριθμός των παρατηρήσεων, RSS είναι το άθροισμα τετραγώνων των καταλοίπων (Residual Sum of Squares), k' είναι ο αριθμός των επεξηγηματικών μεταβλητών (συν τον σταθερό όρο).

Ο πρώτος όρος $nln(RSS/n)$ εκφράζει την ποιότητα προσαρμογής του μοντέλου (μικρότερο RSS οδηγεί σε μικρότερο AIC), ενώ ο δεύτερος όρος $2k$ λειτουργεί ως όρος ποινής (penalty term) που αυξάνεται γραμμικά με τον αριθμό των παραμέτρων, αποθαρρύνοντας την άσκοπη πολυπλοκότητα. Το μοντέλο με τη χαμηλότερη τιμή AIC θεωρείται το βέλτιστο [11].

Η μέθοδος που ακολουθείται είναι η πλήρης διερεύνηση (Full Enumeration), όπου υπολογίζεται το AIC για καθένα από τα 8.192 πιθανά μοντέλα και επιλέγεται εκείνο με την ελάχιστη τιμή. Αυτή η προσέγγιση, αν και υπολογιστικά απαιτητική για μεγάλο p, εγγυάται την εύρεση του global βέλτιστου μοντέλου σύμφωνα με το επιλεγμένο κριτήριο.

Για την εύρεση του βέλτιστου μοντέλου, αναπτύχθηκε ο παρακάτω κώδικας που υλοποιεί την πλήρη διερεύνηση του χώρου των μοντέλων. Αρχικά φορτώνεται το σύνολο δεδομένων `Boston` από το πακέτο `MASS` και ορίζεται η μεταβλητή απόκρισης `medv`, καθώς και το σύνολο των $p = 13$ επεξηγηματικών μεταβλητών. Δημιουργήθηκε η λίστα `all_model_subsets` που περιέχει όλα τα δυνατά υποσύνολα των 13 επεξηγηματικών

μεταβλητών, από το κενό σύνολο (μοντέλο μόνο με σταθερό όρο) μέχρι το πλήρες σύνολο (μοντέλο με όλες τις μεταβλητές). Με τη χρήση ενός βρόγχου (loop), για κάθε υποσύνολο μεταβλητών: κατασκευάζεται ο τύπος του αντίστοιχου γραμμικού μοντέλου (`model_formula_string`), προσαρμόζεται στα δεδομένα με τη συνάρτηση `lm()` και υπολογίζεται η τιμή AIC του μοντέλου με τη συνάρτηση `AIC()`. Αν το νέο AIC είναι μικρότερο από το τρέχον καλύτερο (`best_model_aic_value`), ενημερώνονται τα `best_model_aic_value` και `best_model_formula_string`. Τελικά, επιστρέφεται το μοντέλο M1 με το ελάχιστο AIC.

```r
# Exhaustive model selection using AIC
# Load the dataset
library(MASS)
data(Boston)

# Define response and predictor variables
response <- "medv"
predictors <- names(Boston)[names(Boston) != response]
p <- length(predictors)

# Generate all possible subsets of predictors (2^p total)
all_subsets <- list()
for (k in 0:p) {
  subsets_k <- combn(predictors, k, simplify = FALSE)
  all_subsets <- c(all_subsets, subsets_k)}

# Initialize variables to track the best model
best_aic <- Inf
best_formula <- NULL

# Loop through all predictor subsets
for (vars in all_subsets) {
  # If subset is empty, model includes only the intercept
  if (length(vars) == 0) {
    formula_str <- paste(response, "~ 1")
  } else {
    formula_str <- paste(response, "~", paste(vars, collapse = " + "))  }

  # Fit the model and compute AIC
  current_model <- lm(as.formula(formula_str), data = Boston)
  current_aic <- AIC(current_model)

  # Update best model if current AIC is lower
  if (current_aic < best_aic) {
    best_aic <- current_aic
    best_formula <- formula_str}}

# Fit the final best model (M1)
M1 <- lm(as.formula(best_formula), data = Boston)

# Print the results
cat("Model M1 (Minimum AIC)\n")
cat("Model formula:", best_formula, "\n")
cat("Minimum AIC:", round(best_aic, 3), "\n")
```

Μετά την αξιολόγηση των $2^{13} = 8.192$ πιθανών υπομοντέλων, λαμβάνεται το μοντέλο (M1):

`medv` $\sim$ `crim + zn + chas + nox + rm + dis + rad + tax + ptratio + black + lstat`,

το οποίο ελαχιστοποιεί το κριτήριο AIC. Η ελάχιστη τιμή AIC είναι $AIC_min = 3023.726$.

Το μοντέλο M1 περιλαμβάνει 11 από τις 13 διαθέσιμες επεξηγηματικές μεταβλητές. Οι μεταβλητές `age` και `indus` δεν συμπεριλήφθηκαν, γεγονός που υποδηλώνει ότι —υπό το πρίσμα του AIC— η προσθήκη τους δεν βελτιώνει τη συνολική ισορροπία μεταξύ προσαρμογής και πολυπλοκότητας.

Αξίζει να σημειωθεί ότι η πλήρης εξερεύνηση, παρότι υπολογιστικά δαπανηρή, εξασφαλίζει την ολική βελτιστοποίηση του κριτηρίου AIC, χωρίς τον κίνδυνο εγκλωβισμού σε τοπικά ελάχιστα, όπως μπορεί να συμβεί με ευρετικές μεθόδους.

Ερώτημα 4 (β) - Επιλογή Μεταβλητών μέσω Lasso και Cross-Validation

Το πρόβλημα της επιλογής μεταβλητών μπορεί να προσεγγιστεί και με τη μέθοδο Lasso (Least Absolute Shrinkage and Selection Operator).

Η μέθοδος Lasso εισάγει μια ποινή στο άθροισμα των απόλυτων τιμών των συντελεστών, οδηγώντας αρκετούς από αυτούς σε ακριβώς μηδενικές τιμές, και επομένως εκτελεί ταυτόχρονα συστολή (shrinkage) των συντελεστών και αυτόματη επιλογή μεταβλητών (variable selection). Η εκτίμηση των συντελεστών $\boldsymbol{\beta}$ γίνεται ως λύση του εξής προβλήματος:

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

όπου ο πρώτος όρος είναι το κλασικό Άθροισμα Τετραγώνων των Υπολοίπων (RSS) και ο δεύτερος όρος είναι η ποινή L1 (L1 penalty), η οποία είναι ανάλογη του αθροίσματος των απόλυτων τιμών των συντελεστών. Η παράμετρος λ (lambda) είναι η παράμετρος κανονικοποίησης που ελέγχει την ισχύ της ποινής.

Η ιδιαιτερότητα της ποινής L1 είναι ότι, καθώς η τιμή του λ αυξάνεται, όχι μόνο συρρικνώνει τους συντελεστές προς το μηδέν (όπως η Ridge regression), αλλά μπορεί να τους μηδενίσει εντελώς. Αυτό έχει ως αποτέλεσμα τον αποκλεισμό των αντίστοιχων μεταβλητών από το μοντέλο, πραγματοποιώντας έτσι επιλογή μεταβλητών [10, 3].

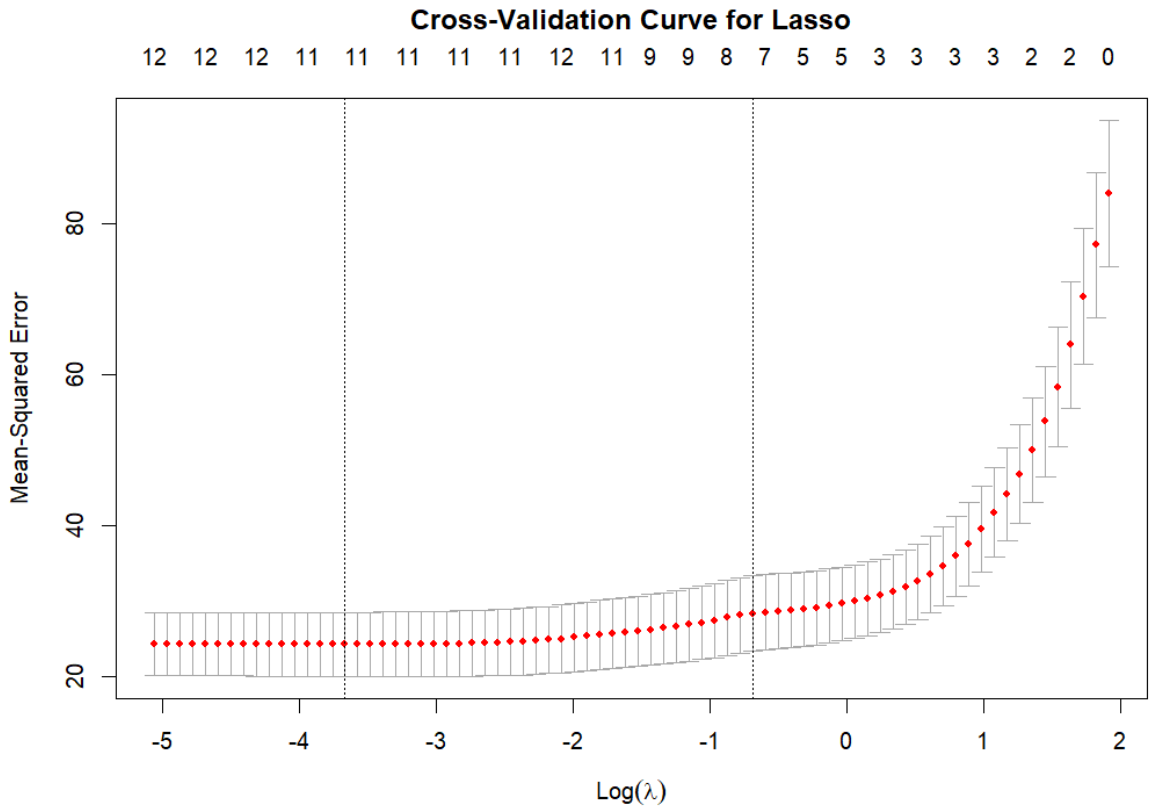
Η επιλογή της βέλτιστης τιμής για το λ είναι κρίσιμη. Μια πολύ μικρή τιμή λ οδηγεί σε ένα μοντέλο παρόμοιο με την απλή γραμμική παλινδρόμηση, ενώ μια πολύ μεγάλη τιμή θα μηδενίσει όλους τους συντελεστές. Για την εύρεση του βέλτιστου λ, θα χρησιμοποιηθεί η μέθοδος της k-fold Διασταυρούμενης Επικύρωσης (k-fold Cross-Validation). Η διαδικασία περιλαμβάνει:

1. Τον διαχωρισμό των δεδομένων σε k (συνήθως 10) ισομεγέθη υποσύνολα (folds).
2. Για κάθε τιμή λ σε ένα εύρος τιμών, εκπαιδεύεται k φορές το μοντέλο Lasso, χρησιμοποιώντας κάθε φορά k-1 υποσύνολα για εκπαίδευση και το ένα που απομένει για επικύρωση.
3. Υπολογίζεται το μέσο τετραγωνικό σφάλμα πρόβλεψης (MSE) στο σύνολο επικύρωσης.
4. Το τελικό CV-MSE για κάθε λ είναι ο μέσος όρος των k MSE που υπολογίστηκαν.

Σύμφωνα με το ζητούμενο, επιλέγεται το λ που ακολουθεί τον κανόνα της μίας τυπικής απόκλισης (one-standard-error rule). Αντί να επιλέξουμε το λ που δίνει το απόλυτο ελάχιστο CV-MSE (lambda.min), επιλέγουμε τη μεγαλύτερη τιμή λ (δηλαδή το πιο απλό/οικονομικό μοντέλο) για την οποία το CV-MSE βρίσκεται εντός μίας τυπικής απόκλισης από το ελάχιστο. Αυτή η προσέγγιση προτιμάται συχνά γιατί οδηγεί σε πιο λιτά μοντέλα με στατιστικά ισοδύναμη προγνωστική ικανότητα [3, Section 3.4.3].

Η ανάλυση υλοποιήθηκε με τη χρήση του πακέτου `glmnet` της R. Αρχικά, δημιουργήθηκε ο πίνακας των επεξηγηματικών μεταβλητών `X` και το διάνυσμα της μεταβλητής απόκρισης `y`. Χρησιμοποίηθηκε η συνάρτηση `cv.glmnet()` για να εκτελεστεί η παλινδρόμηση Lasso με ταυτόχρονη 10-fold διασταυρούμενη επικύρωση (`alpha=1` ορίζει τη Lasso). Η επιλογή `standardize=TRUE` εξασφαλίζει ότι οι μεταβλητές τυποποιούνται πριν την προσαρμογή, κάτι που είναι απαραίτητο σε ποινικοποιημένες μεθόδους. Από το αντικείμενο που επιστρέφεται, εξάγονται οι τιμές `lambda.min` (το λ που ελαχιστοποιεί το CV-MSE) και `lambda.1se` (το λ που προκύπτει από τον κανόνα της μίας τυπικής απόκλισης). Το τελικό μοντέλο Lasso προσαρμόζεται καλώντας ξανά τη συνάρτηση `glmnet()`, αλλά αυτή τη φορά παρέχοντας ως τιμή για την παράμετρο `lambda` την `lambda.1se`. Τέλος, εξάγονται και παρουσιάζονται οι συντελεστές του τελικού μοντέλου για να αναγνωριστούν οι μεταβλητές που παρέμειναν στο μοντέλο (εκείνες με μη μηδενικό συντελεστή).

```r
# Lasso Regression using glmnet and Cross-Validation

# Load required libraries
library(glmnet)
library(MASS)

# Define response and predictor matrix
y <- Boston$medv
X <- as.matrix(Boston[, names(Boston) != "medv"])

# Fit Lasso model with cross-validation (10-fold by default)
set.seed(123)
cv_lasso <- cv.glmnet(X, y, alpha = 1, standardize = TRUE)

# Plot the cross-validation curve (optional)
plot(cv_lasso)
title("Cross-Validation Curve for Lasso", line = 2.5)

# Lambda that minimizes CV error
lambda_min <- cv_lasso$lambda.min

# Lambda using 1-standard-error rule (simpler model)
lambda_1se <- cv_lasso$lambda.1se

# Fit final model using lambda_1se
lasso_model <- glmnet(X, y, alpha = 1, lambda = lambda_1se, standardize = TRUE)

# Extract coefficients
lasso_coefs <- coef(lasso_model)

# Print results
cat("Lasso Regression Results\n")
cat("Lambda (min):", round(lambda_min, 5), "\n")
cat("Lambda (1-SE rule):", round(lambda_1se, 5), "\n\n")

cat("Selected predictors (non-zero coefficients):\n")
selected <- rownames(lasso_coefs)[lasso_coefs[, 1] != 0]
print(selected[-1])  # exclude intercept

# print(lasso_coefs) # to see all coefficients
```



Σχήμα 10. *Καμπύλη διασταυρούμενης επικύρωσης για την επιλογή της παραμέτρου λ στο μοντέλο Lasso.*

Από τη διαδικασία της διασταυρούμενης επικύρωσης λαμβάνεται το διάγραμμα που παρατίθεται στο Σχήμα 10, το οποίο δείχνει το Μέσο Τετραγωνικό Σφάλμα (MSE) ως συνάρτηση του λογαρίθμου της παραμέτρου ποινής λ, Log(λ). Τα κόκκινα σημεία δείχνουν το μέσο CV-MSE για κάθε τιμή λ, ενώ οι γκρι μπάρες το διάστημα εμπιστοσύνης μίας τυπικής απόκλισης. Ο άξονας στο πάνω μέρος του γραφήματος δείχνει τον αριθμό των μη μηδενικών συντελεστών (δηλαδή των επιλεγμένων μεταβλητών) για κάθε τιμή του λ. Όπως αναμένεται, καθώς το λ αυξάνεται (κίνηση από αριστερά προς τα δεξιά), ο αριθμός των μεταβλητών μειώνεται.

Οι τιμές για τις παραμέτρους λ που προέκυψαν είναι:

- `lambda.min = 0.02552`: Η τιμή λ που ελαχιστοποιεί το σφάλμα διασταυρούμενης επικύρωσης.
- `lambda.1se = 0.50092` : Η τιμή λ που προκύπτει από τον κανόνα της μίας τυπικής απόκλισης.

Ακολουθώντας την εκφώνηση, επιλέγεται το `lambda.1se`, οπότε το τελικό μοντέλο περιλαμβάνει τις ακόλουθες 7 επεξηγηματικές μεταβλητές: `"crim" "chas" "rm" "dis" "ptratio" "black" "lstat"`.

Συμπερασματικά, η μεθοδολογία Lasso με διασταυρούμενη επικύρωση και εφαρμογή του κανόνα της μίας τυπικής απόκλισης, επέλεξε ένα πιο λιτό (parsimonious) μοντέλο σε σύγκριση με το μοντέλο M1 (11 μεταβλητές) που προέκυψε από την πλήρη διερεύνηση με κριτήριο το AIC. Αυτό αναδεικνύει μια βασική ιδιότητα της Lasso, η οποία, σε συνδυασμό με τον κανόνα 1-SE, τείνει να παράγει απλούστερα μοντέλα, μηδενίζοντας συντελεστές που προσφέρουν μικρή προγνωστική συνεισφορά. Το τελικό μοντέλο με τις 7 μεταβλητές θεωρείται στατιστικά ισοδύναμο σε προγνωστική ισχύ με το πιο πολύπλοκο μοντέλο που αντιστοιχεί στο lambda.min, αλλά είναι προτιμότερο λόγω της απλότητάς του.

Ερώτημα 4 (γ) - Εκτίμηση Διαστήματος Εμπιστοσύνης μέσω Residual Bootstrap

Για το υπομοντέλο M1 που επιλέχθηκε στο ερώτημα (α) βάσει του κριτηρίου AIC, εφαρμόζεται η μέθοδος Residual Bootstrap για την κατασκευή ενός 95% διαστήματος εμπιστοσύνης για τον συντελεστή της επεξηγηματικής μεταβλητής `rm` (μέσος αριθμός δωματίων ανά κατοικία).

Η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη όταν οι υποθέσεις της κλασικής γραμμικής παλινδρόμησης (όπως η κανονικότητα και η ομοσκεδαστικότητα των σφαλμάτων) ενδέχεται να μην ισχύουν [5]. Η βασική ιδέα της Residual Bootstrap είναι η εξής:

Η μέθοδος Bootstrap είναι μια τεχνική επαναδειγματοληψίας που χρησιμοποιείται για την εκτίμηση της κατανομής δειγματοληψίας ενός στατιστικού μεγέθους, καθώς και για την κατασκευή διαστημάτων εμπιστοσύνης και τον έλεγχο υποθέσεων. Είναι ιδιαίτερα χρήσιμη όταν οι υποθέσεις της κλασικής γραμμικής παλινδρόμησης (όπως η κανονικότητα και η ομοσκεδαστικότητα των σφαλμάτων) ενδέχεται να μην ισχύουν. Η βασική ιδέα της Residual Bootstrap είναι η εξής:

1. **Προσαρμογή αρχικού μοντέλου παλινδρόμησης**: Αρχικά, προσαρμόζεται το γραμμικό μοντέλο M_1 στα αρχικά δεδομένα $(\mathbf{y}, \mathbf{X})$ και λαμβάνονται οι προβλεπόμενες τιμές ($\hat{\mathbf{y}}$) και τα υπόλοιπα ($\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$). Το μοντέλο είναι:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

   Οι εκτιμήσεις είναι $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ και $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.

2. **Δημιουργία B δειγμάτων Bootstrap**: Θεωρούμε τα υπολογισμένα υπόλοιπα $\{e_1, e_2, \ldots, e_n\}$ ως μια εμπειρική κατανομή που προσεγγίζει την άγνωστη κατανομή των πραγματικών σφαλμάτων ϵ. Στη συνέχεια, δημιουργούμε B νέα "ψευδο-δείγματα" (bootstrap samples) της μεταβλητής απόκρισης. Κάθε νέο δείγμα $\mathbf{y}^{*(b)}$ κατασκευάζεται ως εξής:

$$y_i^{*(b)} = \hat{y}_i + e_i^{*(b)}, \quad \text{για } b = 1, \ldots, B$$

   όπου $\mathbf{e}^{*(b)}$ είναι ένα διάνυσμα υπολοίπων που προκύπτει από δειγματοληψία με επανατοποθέτηση από το αρχικό σύνολο των υπολοίπων $\{e_i\}$. Με τον τρόπο αυτό, διατηρείται σταθερή η δομή του μοντέλου (που εκφράζεται από τις $\hat{\mathbf{y}}$), ενώ προσομοιώνουμε τη στοχαστικότητα των δεδομένων μέσω των υπολοίπων επαναδειγματοληψίας.

3. **Επανεκτίμηση του μοντέλου**: Για καθένα από τα B δείγματα $\mathbf{y}^{*(b)}$, προσαρμόζεται ξανά το ίδιο γραμμικό μοντέλο (με τις ίδιες επεξηγηματικές μεταβλητές $\mathbf{X}$) και εκτιμώνται οι συντελεστές $\hat{\boldsymbol{\beta}}^{*(b)}$. Αποθηκεύεται η εκτίμηση του συντελεστή που μας ενδιαφέρει, δηλαδή του $\hat{\beta}_{\text{rm}}^{*(b)}$.

4. **Κατασκευή Δ.Ε.**: Μετά την ολοκλήρωση των B επαναλήψεων, λαμβάνεται μια κατανομή bootstrap των εκτιμήσεων του συντελεστή $\left\{\hat{\beta}_{\text{rm}}^{*(1)}, \hat{\beta}_{\text{rm}}^{*(2)}, \ldots, \hat{\beta}_{\text{rm}}^{*(B)}\right\}$. Αυτή η κατανομή προσεγγίζει τη δειγματική κατανομή του εκτιμητή $\hat{\beta}_{\text{rm}}$. Για την κατασκευή διαστημάτων εμπιστοσύνης βασισμένων στα ποσοστιαία σημεία (percentile confidence intervals) ταξινομούνται οι B εκτιμήσεις του συντελεστή σε αύξουσα σειρά και επιλέγονται τα κατάλληλα ποσοστιαία σημεία. Για ένα $(1 - \alpha)$ διάστημα εμπιστοσύνης χρησιμοποιούνται τα ποσοστιαία σημεία:

$$\left(\frac{\alpha}{2},\ 1 - \frac{\alpha}{2}\right)$$

   της εμπειρικής κατανομής των B εκτιμήσεων από το Bootstrap.

Για την υλοποίηση του Residual Bootstrap στην `R`, αρχικά ορίζεται ο αριθμός των επαναλήψεων Bootstrap: `B = 2000` και από το μοντέλο `M1` εξάγονται οι προβλεπόμενες τιμές (`fitted_vals`) και τα υπόλοιπα

(`residuals_original`). Κατόπιν, εκτελείται B φορές ένας βρόγχος, σε κάθε επανάληψη του οποίου γίνεται επαναδειγματοληψία με επανατοποθέτηση από τα `residuals_original` για τη δημιουργία των `resampled_residuals`. Δημιουργούνται οι νέες τιμές απόκρισης `y_star` προσθέτοντας τα `resampled_residuals` στις αρχικές `fitted_vals`, καθώς επίσης και ένα νέο σύνολο δεδομένων `data_star` αντικαθιστώντας την αρχική μεταβλητή απόκρισης `medv` με τις τιμές `y_star`. Το μοντέλο `M1` (με την ίδια δομή επεξηγηματικών μεταβλητών) προσαρμόζεται στα νέα δεδομένα `data_star` και ο συντελεστής που αντιστοιχεί στη μεταβλητή `rm` από το `model_star` αποθηκεύεται στο `bootstrap_coefs`. Μετά το τέλος του βρόγχου, υπολογίζονται τα 2.5% (`ci_low`) και τα 97.5% (`ci_high`) ποσοστιαία σημεία από τις B τιμές του `bootstrap_coefs`, χρησιμοποιώντας τη συνάρτηση `quantile()`. Αυτά τα όρια ορίζουν το **95% διάστημα εμπιστοσύνης** βασισμένο στα ποσοστιαία σημεία της εμπειρικής κατανομής bootstrap.

```r
# Residual Bootstrap for 95% CI of the coefficient of 'rm'

# Number of bootstrap replications
B <- 2000

# Extract fitted values and residuals from M1
fitted_vals <- fitted(M1)
residuals_original <- resid(M1)

# Create vector to store bootstrap estimates of the coefficient of 'rm'
bootstrap_coefs <- numeric(B)

# Residual bootstrap loop
set.seed(123)
for (b in 1:B) {
  # Resample residuals with replacement
  resampled_residuals <- sample(residuals_original, replace = TRUE)

  # Generate new response values: y* = fitted + resampled residuals
  y_star <- fitted_vals + resampled_residuals

  # Fit the same model to the new response
  data_star <- Boston
  data_star$medv <- y_star

  model_star <- lm(formula(M1), data = data_star)

  # Extract the coefficient for 'rm'
  bootstrap_coefs[b] <- coef(model_star)["rm"]
}

# Compute the 95% percentile bootstrap confidence interval
ci_low <- quantile(bootstrap_coefs, 0.025)
ci_high <- quantile(bootstrap_coefs, 0.975)

# Print results
cat("--- 95% Bootstrap CI for coefficient of 'rm' ---\n")
cat("Lower bound (2.5%):", round(ci_low, 4), "\n")
cat("Upper bound (97.5%):", round(ci_high, 4), "\n")
```

Από την εφαρμογή της μεθόδου Residual Bootstrap με 2000 επαναλήψεις για τον συντελεστή της μεταβλητής `rm` στο μοντέλο `M1`, ελήφθη το 95% διάστημα εμπιστοσύνης:

95% CI: [2.9763, 4.5481]

Παρατηρείται ότι το διάστημα εμπιστοσύνης δεν περιέχει την τιμή μηδέν (0). Αυτό αποτελεί ισχυρή ένδειξη ότι ο συντελεστής της μεταβλητής `rm` είναι στατιστικά σημαντικός στο επίπεδο σημαντικότητας α=5%. Επομένως, ο μέσος αριθμός δωματίων (`rm`) έχει στατιστικά σημαντική επίδραση στη μέση αξία ιδιοκατοίκησης (`medv`).

Δεδομένου, επίσης, ότι ολόκληρο το διάστημα εμπιστοσύνης αποτελείται από θετικές τιμές, μπορούμε να συμπεράνουμε με βεβαιότητα 95% ότι υπάρχει μια θετική σχέση μεταξύ του αριθμού των δωματίων και της αξίας της κατοικίας. Αυτό σημαίνει ότι, διατηρώντας σταθερές τις υπόλοιπες μεταβλητές του μοντέλου, μια αύξηση στον αριθμό των δωματίων αναμένεται να οδηγήσει σε αύξηση της μέσης αξίας της κατοικίας. Συγκεκριμένα, για κάθε επιπλέον δωμάτιο, η αξία της κατοικίας αναμένεται να αυξηθεί κατά μια τιμή που, με 95% βεβαιότητα, κυμαίνεται μεταξύ 2.976 και 4.548 (σε μονάδες $1000).

Βιβλιογραφία

[1] Δημήτριος Φουσκάκης. *Density Estimation.* Σημειώσεις Μαθήματος "Υπολογιστική Στατιστική & Στοχαστική Βελτιστοποίηση". 2024. URL: http://www.math.ntua.gr/%7Efouskakis/Computational_Stats/Slides/1.density_estimation.pdf (επίσκεψη 19/06/2024).

[2] Δημήτριος Φουσκάκης. *Cross-Validation.* Σημειώσεις Μαθήματος "Υπολογιστική Στατιστική & Στοχαστική Βελτιστοποίηση". 2024. URL: http://www.math.ntua.gr/%7Efouskakis/Computational_Stats/Slides/4.cross_validation.pdf (επίσκεψη 19/06/2024).

[3] Trevor Hastie, Robert Tibshirani και Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd. New York: Springer Science & Business Media, 2009. ISBN: 978-0-387-84857-0.

[4] R Core Team. *Kernel Regression Smoother (Function: ksmooth).* R Documentation. 2024. URL: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/ksmooth.html (επίσκεψη 19/06/2024).

[5] Δημήτριος Φουσκάκης. *Resampling Methods: Jackknife-Bootstrap.* Σημειώσεις Μαθήματος "Υπολογιστική Στατιστική & Στοχαστική Βελτιστοποίηση". 2024. URL: http://www.math.ntua.gr/%7Efouskakis/Computational_Stats/Slides/3.resampling.pdf (επίσκεψη 19/06/2024).

[6] Δημήτριος Φουσκάκης. *Stochastic Simulation.* Σημειώσεις Μαθήματος "Υπολογιστική Στατιστική & Στοχαστική Βελτιστοποίηση". 2024. URL: http://www.math.ntua.gr/%7Efouskakis/Computational_Stats/Slides/2.stochastic_simulation.pdf (επίσκεψη 19/06/2024).

[7] Christian P. Robert και George Casella. *Monte Carlo Statistical Methods.* 2nd. New York: Springer, 2004. ISBN: 978-0-387-21239-5.

[8] Δημήτριος Φουσκάκης. *Markov Chain Monte Carlo (MCMC).* Σημειώσεις Μαθήματος "Υπολογιστική Στατιστική & Στοχαστική Βελτιστοποίηση". 2024. URL: http://www.math.ntua.gr/%7Efouskakis/Computational_Stats/Slides/5.MCMC.pdf (επίσκεψη 19/06/2024).

[9] Δημήτριος Φουσκάκης. *Expectation-Maximization Algorithm.* Σημειώσεις Μαθήματος "Υπολογιστική Στατιστική & Στοχαστική Βελτιστοποίηση". 2024. URL: http://www.math.ntua.gr/%7Efouskakis/Computational_Stats/Slides/7.EM.pdf (επίσκεψη 19/06/2024).

[10] Δημήτριος Φουσκάκης. *Variable Selection.* Σημειώσεις Μαθήματος "Υπολογιστική Στατιστική & Στοχαστική Βελτιστοποίηση". 2024. URL: http://www.math.ntua.gr/%7Efouskakis/Computational_Stats/Slides/8.variable_selection_Lasso.pdf (επίσκεψη 19/06/2024).

[11] Δημήτριος Φουσκάκης. *Stochastic Optimisation Methods.* Σημειώσεις Μαθήματος "Υπολογιστική Στατιστική & Στοχαστική Βελτιστοποίηση". 2024. URL: http://www.math.ntua.gr/%7Efouskakis/Computational_Stats/Slides/6.stochastic_optimisation.pdf (επίσκεψη 19/06/2024).