Assignment Part II

Silpa Soni Nallacheruvu (19980824-5287), Elva Wallimann (19780306-T063)

2024-10-14

set.seed(900101)

Summary

The Assignment focuses on applying knowledge related to Wald Statistics, Generalized likelihood ratio statistics, Profile likelihood and Logistic Regression using the Newton-Raphson's Algorithm from previous assignment. The first task is about verifying the accuracy of the Wald Statistics against the built-in R's results of the same z values. The second task is about verifying the accuracy of the Generalized likelihood ratio statistics and computing the p-value to confirm the significance of model parameters. The third task is about testing the different model parameters against the null hypothesis of zero, except "Alder" and "Utbildare". The fourth task is about in-depth analysis of the parameter θ_{Kon} based on a grid of possible values, computing the confidence intervals and plotting the profile and estimated likelihood.

Task 1

Approach :

To calculate the Wald statistics from the ML estimates, the formula

$$\frac{\hat{\theta}-\theta_0}{\sqrt{\mathrm{diag}(I(\hat{\theta})^{-1})}}$$

was applied where I($\hat{\theta}$) represents the Fisher Information Matrix and $\sqrt{\mathrm{diag}(I(\hat{\theta})^{-1})}$ represents the standard error. These values were then compared with the z values from R in the provided summary.

Code :

```r
# ---- Task_1 ----

#Verify if z values are Wald statistics using I and NR functions

# Wald statistic = (theta_estimate - theta0)/(standard_error)
#Consider the theta estimate after 3 iterations as per Part-I
theta0 = c(0, 0, 0, 0)
theta_estimate <- NR(theta0, 3, y, X)
I_matrix <- I(theta_estimate, X)
standard_error <- sqrt(diag(solve(I_matrix)))
wald_statistic <- theta_estimate/standard_error

#Compare the wald estimate output with z values in summary
z_values <- summary(modell)$coefficients[, "z value"]
```

```
# Creating a comparison data frame
comparison_z_values <- data.frame(
  "Z_values_R_model" = z_values,
  "Z_values_computed" = wald_statistic
  )
```

Output :

A comparison table between the z values in the provided R summary and the z values from ML estimates using the Wald statistics are attached for further reference.

```
comparison_z_values
```

```
##                           Z_values_R_model Z_values_computed
## (Intercept)                      0.5354616         0.5354605
## Alder                           -4.0833097        -4.0833063
## KonMan                           2.9342844         2.9342828
## UtbildareTrafikskola             6.4608428         6.4608407
```

Observation :

The z values derived from the ML Estimates, computed using the Wald Statistics align with the standard errors provided in the R summary. The values are consistent up until the 5th decimal place. This suggests that the Wald Statistics are reliable and since Wald Statistics, which is defined as : $\frac{(\hat{\theta}-\theta)}{\text{se}(\hat{\theta})}$, follows an asymptotic standard normal distribution under H_0, and so, it can be used to derive accurate z values.

Task 2

Approach :

To calculate the Generalized likelihood ratio statistics from the ML estimates, the formula

$2{*}\log(\frac{L_p(\hat{\theta}_{\text{ML}})}{L_p(\hat{\theta}_0)})$

was applied, where $\{L_p(\hat{\theta}_{\text{ML}})\}$ represents the Profile Likelihood at ML estimate of theta parameter (computed from task1) and $\{L_p(\hat{\theta}_0)\}$ represents the Profile Likelihood at initial theta of the null hypothesis. These values were then compared with the squared values of wald statistics from task1 to verify the magnitude of the Generalized likelihood ratio. the Generalized likelihood ratio statistic.

Code :

```
# ---- Task_2 ----

# Compute generalized likelihood ratio statistics that corresponds
# to Wald Statistics in Task1
# GN_L = 2 x log(profile_likelihood(theta_estimate)/profile_likelihood(theta0))

# Generalized likelihood ratio statistics
compute_gnl_ratio <- function(XP) {
  #compute profile log likelihood for theta estimate
  profile_theta0 <- rep(0, ncol(XP))
  profile_theta_estimate0 <- NR(profile_theta0, 3, y, XP)
  profile_log_likelihood_estimate0 <- l(profile_theta_estimate0, y, XP)
  # using the theta_estimate from task1 generated for Wald Statistics here
  # since it contains the MLEs for theta and eta
  log_likelihood_estimate <- l(theta_estimate, y, X)

  gnl_ratio <- 2* (log_likelihood_estimate - profile_log_likelihood_estimate0)
  return(gnl_ratio)
}

#compute the gnl statistics ratio for each parameter and profile the X matrix
# by each parameter
gnl_ratio_intercept <- compute_gnl_ratio(X[, -1])
gnl_ratio_alder <- compute_gnl_ratio(X[, -2])
gnl_ratio_kon <- compute_gnl_ratio(X[ ,-3])
gnl_ratio_utbildare <- compute_gnl_ratio(X[, -4])

# Creating a comparison data frame for gnl ratio
comparison_gnl_values <- data.frame(
  "squared_wald_statistic" = wald_statistic^2,
  "generalized_likelihood_ratio_statistics" =
    c(gnl_ratio_intercept, gnl_ratio_alder, gnl_ratio_kon, gnl_ratio_utbildare)
)

# Determine the corresponding P-values
p_value_intercept <- pchisq(gnl_ratio_intercept, df = 1, lower.tail = FALSE)
p_value_alder <- pchisq(gnl_ratio_alder, df = 1, lower.tail = FALSE)
p_value_kon <- pchisq(gnl_ratio_kon, df = 1, lower.tail = FALSE)
p_value_utbildare <- pchisq(gnl_ratio_utbildare, df = 1, lower.tail = FALSE)

p_values <- data.frame(
  "intercept_profile" = p_value_intercept,
  "alder_profile" = p_value_alder,
  "kon_profile" = p_value_kon,
  "utbildare_profile" = p_value_utbildare)

#Compare the p values output with p values in summary
p_values_summary <- summary(modell)$coefficients[, "Pr(>|z|)"]

# Creating a comparison data frame
comparison_p_values <- data.frame(
  "P_values_R_model" = p_values_summary,
  "P_values_computed" = c(p_value_intercept, p_value_alder, p_value_kon, p_value_utbildare)
)
```

Output :

A comparison table between the squared wald statistics values generated in the previous task and the generalized likelihood ratio values for each parameter are attached for further reference. in the previous task and the generalized likelihood ratio statistic values for each parameter are attached for further reference.

```
comparison_gnl_values
```

```
##                         squared_wald_statistic
## (Intercept)                          0.2867179
## Alder                               16.6733903
## KonMan                            8.6100156
## UtbildareTrafikskola             41.7424629
##                        generalized_likelihood_ratio_statistics
## (Intercept)                                          0.2868917
## Alder                                               17.3154264
## KonMan                                               8.7213578
## UtbildareTrafikskola                                43.1541479
```

A list of p values generated from generalised likelihood ratio values A list of p values generated from generalised likelihood ratio statistic values for each parameter are attached for further reference.

```
p_values
```

```
##   intercept_profile alder_profile kon_profile utbildare_profile
## 1         0.5922194  3.166062e-05 0.003145037      5.059253e-11
```

A comparison table between the p values generated from generalised likelihood ratio statistic value and the R summary for each parameter are attached for further reference.

```
comparison_p_values
```

```
##                        P_values_R_model P_values_computed
## (Intercept)                5.923307e-01      5.922194e-01
## Alder                      4.439878e-05      3.166062e-05
## KonMan                     3.343177e-03      3.145037e-03
## UtbildareTrafikskola       1.041214e-10      5.059253e-11
```

Observation :

1. The Generalized likelihood ratio statistic values derived from the ML Estimates, computed using the profile log-likelihoods align in magnitude with the squared Wald Statistics. The values are close in nature. This is because the Wald Statistics follow an asymptotic $N(0, 1)$ under H_0 and generalized likelihood ratio follows the asymptotic $\chi^2(1)$ distribution. Hence, the generalized likelihood ratio values are of the same order of magnitude as the squared Wald Statistics. $N(0, 1)$ under H_0 and generalized likelihood ratio statistic follows the asymptotic $\chi^2(1)$ distribution. Hence, the generalized likelihood ratio statistic values are of the same order of magnitude as the squared Wald Statistics.
2. The p-values for profile likelihoods of kon, alder and utbildare lie in the extreme tails, which is less than 0.005. Hence, we can reject H_0 for the parameters kon, alder and utbildare. The p-value of the intercept parameter lies in between 0.005 and 0.995. Hence, we can retain H_0 for the intercept parameter.
3. We can verify the Generalized likelihood ratio statistic values by comparing its p-values with the values from R summary. The comparison is accurate till the 4th decimal place.

Task 3

We first take columns "Alder" and "Utbildare" to form the matrix X_new, and drop the rest (assuming they have parameters of 0 due to H_0: $\theta = \theta_0 = (0, 0)$)

```
X_new = X[, c(2,4) ]
```

Then we estimate the parameter vector η for these two columns.

```
eta <- NR(theta0=c(0, 0), niter=10, y, X_new)
```

Given η and θ_0, we then compute the generalized score statistic:

$T_s(\boldsymbol{\theta}_0) = S(\boldsymbol{\theta}_0, \hat{\boldsymbol{\eta}}_{\boldsymbol{ML}}(\boldsymbol{\theta}_0))^T \, IS(\boldsymbol{\theta}_0, \hat{\boldsymbol{\eta}}_{\boldsymbol{ML}}(\boldsymbol{\theta}_0))^{-1} \, S(\boldsymbol{\theta}_0, \hat{\boldsymbol{\eta}}_{\boldsymbol{ML}}(\boldsymbol{\theta}_0))$.

```
theta_new <- c(0, eta[1], 0, eta[2])
score <- S(theta_new, y, X)
info <- I(theta_new, X)
info_inverse <- solve(info)
Ts_theta0 <- t(score) %*% info_inverse %*% score
```

The corresponding p-value is subsequently calculated

```
df <- length(theta_new)  # Degree of freedom; number of parameters under H_0
p_value <- 1 - pchisq(Ts_theta0, df = df)
p_value
```

```
##            [,1]
## [1,] 0.02803378
```

Task 4

We first decide the suitable grid of values for the parameter θ_{Kon}. From the task 1 we know that the estimated value for θ_{Kon} is 0.4000882. Now we extend this value to a interval of [0.4-1, 0.4+1] with steps of 0.01 as the grid values for θ_{Kon}. Then, we use a for-loop to multiply each of the grid value with the corresponding column in the data X, resulting the offset. Subsequently, we fit a new model with the offset to get $\hat{\eta}_{ML}$ and plug in the grid value of θ_{Kon} to get θ_{new}, which is in turn used to compute the profile likelihood (L_p) using the L function that we built in the last assignment. For comparison, the estimated likelihood (L_e) of θ_{Kon} is also computed in the for loop using the estimated values for elements in η (a result from the task 1) and the grid value of θ_{Kon}.

```
# Grid values for theta_Kon
grid <- seq(0.4-1, 0.4+1, 0.01)

# Compute the profile likelihood
L_p <- numeric(length(grid))
L_e <- numeric(length(grid))

for(i in seq(length(grid))){
  theta_Kon <- grid[i]
  new_model <- glm.fit(x = X[, -3], y = y,
                       offset = theta_Kon * X[, 3],
                       family = binomial())
  theta_new <- c(new_model$coeff[1:2], theta_Kon, new_model$coeff[3])

  # Compute the profile likelihood
  L_p[i] <- L(theta_new, y, X)

  # Compute the estimated likelihood
  L_e[i] <- L(c(theta_estimate[1:2], theta_Kon, theta_estimate[4]), y, X)
}
```

Now we are ready to plot the profile and estimated likelihood for $theta_{Kon}$. To draw the line for the 95% confidence interval, we compute the critical value as the following:

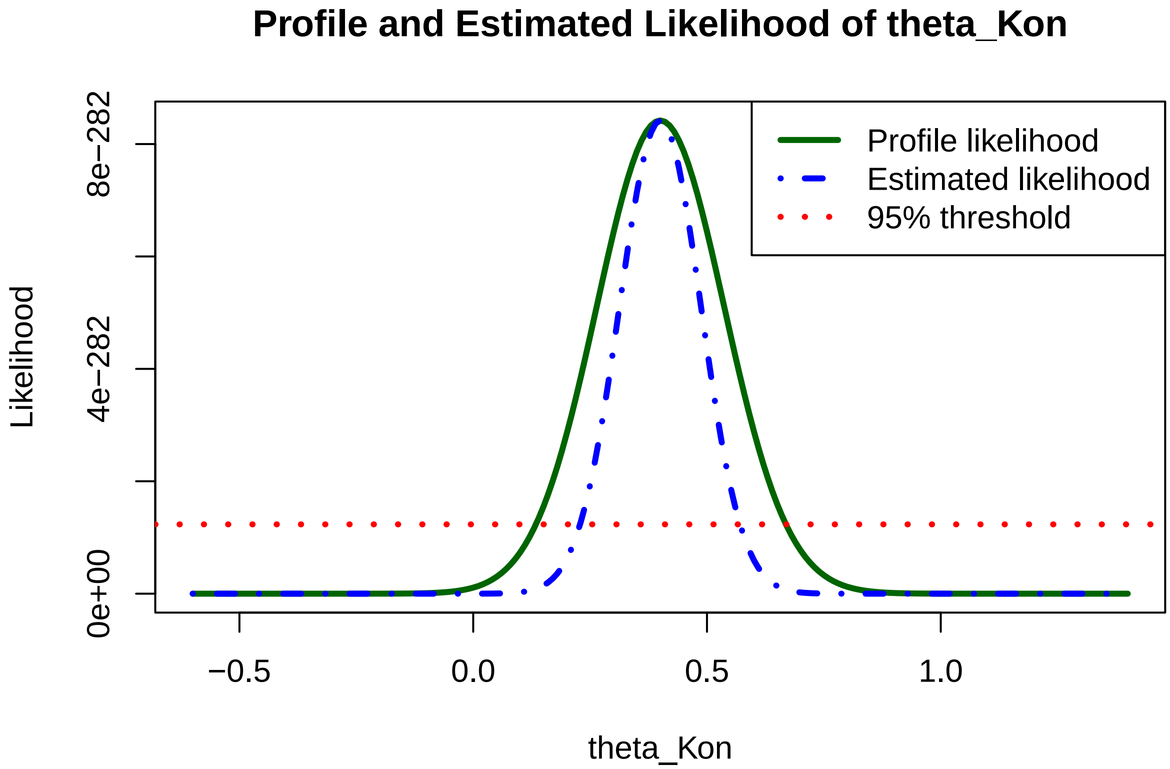
$\max(L_p(\theta_0)) \times \exp\left[-\frac{1}{2}\chi^2_{0.95}(1)\right]$.

```
# Plot the profile likelihood and the estimated likelihood
## Create the plot
plot(grid, L_p, type = "l", col = "darkgreen", lwd=3,
     main = "Profile and Estimated Likelihood of theta_Kon",
     xlab = "theta_Kon",
     ylab = "Likelihood")

## Add the estimated likelihood
lines(grid, L_e, col = "blue", lty = 4, lwd=3)

## Draw a horizontal line for the 95% profile likelihood confidence interval
ci_95 <- max(L_p) * exp(-0.5 * qchisq(0.95, df=1))
abline(h = ci_95, col = "red", lty = 3, lwd=3)

## Add a legend
legend("topright", legend = c("Profile likelihood", "Estimated likelihood", "95% threshold"),
       col = c("darkgreen", "blue", "red"), lty = c(1, 4, 3), lwd = 3)
```



From the plot, we can see that the 95% confidence interval is roughly [0.1, 0.7], which will be compared with the Wald confidence interval in the final step. We first compute the Wald confidence interval using this formula: $\hat{\theta} \pm z_{0.975}\ \mathrm{se}(\hat{\theta})$. The resulting 95% Wald confidence interval is [0.1328479, 0.6673284], which is in line with the plotted 95% interval.

```
se_Kon <- standard_error[3]
theta_Kon_estimated <- theta_estimate[3]
lb <- theta_Kon_estimated - qnorm(0.975) * se_Kon
ub <- theta_Kon_estimated + qnorm(0.975) * se_Kon
cat(sprintf("[%f, %f]", lb, ub))
```

```
## [0.132848, 0.667328]
```