The University of Newcastle

College of Engineering, Science and Environment

School of Information and Physical Sciences

Statistical Inference
STAT2300: Semester 2, 2024

Group Project

Author (s):
Nathan Hill, C3334136
Jacob Saunders, C3262240

October 20, 2024

STAT2300 Group Project

Introduction

Proposal Due: Electronically via Canvas by 11:59pm, Sunday, September 22.

Project Due: Electronically via Canvas by 11:59pm, Sunday, October 20.

Presentation: During the Week 12 lab.

Worth: 30% of your overall grade (20% for the group project and 10% for the group presentation)

- The purpose of this group project is to work collaboratively to apply the statistical concepts and methods of the course to a data set of your group's choosing and present your work to an audience of your peers.
- Please note that your submission should contain your own work. Please refer to the **Student Conduct Rule** for more information.
- The use of generative AI is not permitted for this assessment. **See more details.**

Forming Groups

For this project, you must work in groups of 2 or 3 students. Please formalise your groups in Canvas in the People area. If you're struggling to find someone to work with, there is a thread in the Discussion Board where you can reach out to other students to form groups.

Proposal

The purpose of the proposal is to give you a mechanism to get early feedback on your ideas for the project so that you don't pursue a project that is not viable. The proposal is unweighted, but it will give you an opportunity to outline the plan for your project and me as Course Coordinator the chance to provide guidance or steer you in the right direction as needed. For the proposal, please provide a description of the following:

- **The data:** What data are you going to use? Will you use a real data set or instead performing simulations? If using a real data set, where will you obtain the data? How big will the sample size be? If performing simulations, how many simulations do you intend to perform? What will you be computing in each simulation?
- **The methods you will use to analyse the data:** You must include up to 3 methods we have learned about in STAT2300. What will these methods be able to illuminate about the data?

After your group's proposal is approved, you can proceed with the project.

Project

With the project, you will apply 3 or more concepts from STAT2300 to your chosen data set or simulation design. The project should be a comprehensive written document and it should include the following sections:

1. **Introduction:** In this section, you will provide an overview of the application or theoretical area for your project. Through the methods you will employ, what will you explore about the particular application or theoretical domain you have chosen?
2. **Data or Simulation Design:** In this section, you will provide a full description of the data you will be analysing. More specifically,
    - If you are using a real data set, provide the source of the data set, a description of all of the variables, and the sample size.
    - If you are performing simulations, fully describe the simulation design, including the number of simulations you will be performing and the objects you will create to store the information from the simulations.
3. **Topics Covered:** In this section, you will list 3 or more methods from STAT2300 that you will apply to the data, and what you will use them for.

4. **Results:** In this section, you will apply the methods from the previous section to the data set and provide a full code or working to obtain the results. Note: your work must be completely reproducible. This means:
    - If you make use of any functions contained in add-in packages, you must list these packages and include the appropriate call to library() to load the functions within your code.
    - If you perform any random number generation or random sampling, you must set the seed of the random number generation with the set.seed() function.
5. **Conclusion:** In this section, you will summarise the results and describe what they illustrate about the application or theoretical domain you've mentioned in the introduction.

Group Member Attribution

In addition to the written project, each group must submit a brief report providing details of what each group member contributed to the project. Broadly, this can be aligned to each of the 5 sections above:

1. Introduction
    - Ideas
    - Writing
    - Reviewing
2. Data/Simulation Design
    - Ideas
    - Writing
    - Reviewing
3. Topics Covered
    - Ideas
    - Writing
    - Reviewing
4. Results
    - Writing (text)
    - Writing (code)
    - Reviewing (text and/or code)
5. Conclusion
    - Writing
    - Reviewing

This document must be signed by all group members.

Peer Review

Each member of the group must individually submit a peer reflection form. This gives you the opportunity to anonymously rate and comment on the performance of the other members of your group.

Presentation

In your presentation, your group will have 8 minutes to present your project to the rest of the class during the lab in Week 12. Your presentation should be accompanied by some visual content, such as slides. There are various ways you might format your presentation, but one structure which might work well:

1. **Introduction:** introduce the subject area of your project and why it is interesting. Perhaps briefly indicate the techniques your group applied. ($\sim$2 minutes)
2. **Results:** provide an overview of the results of your project. You don't need to showcase all of the code, but report the main results and how they are meaningful to the subject area. ($\sim$5 minutes)
3. **Conclusion:** provide a summary of your findings and what your group learned about the subject area. ($\sim$1 minute)

1 Introduction

The importance of understanding the diabetes dataset it to further our knowledge of what the key causes of diabetes are, and how they can be predicted. The dataset contains 11 variables, and the target variable is the progression of diabetes. The dataset contains 442 observations and 11 variables and we are understanding the factors involved which are:

- **a**ge - age in years
- **s**ex - male and female
- **b**mi - body mass index
- **b**p - average blood pressure
- **s**1 - TC: total serum cholesterol
- **s**2 - LDL: low-density lipoproteins
- **s**3 - HDL: high-density lipoproteins
- **s**4 - TCH: total cholesterol / HDL
- **s**5 - LTG: possibly log of serum triglycerides level
- **s**6 - GLU: blood sugar level
- **t**arget - measure of disease progression one year after baseline

Each of the feature variables have been mean centered and scaled by the standard deviation times the square root of the number of sample.

Our investigation is to look at the more commonly known variables that affect the progression of diabetes in the observations of this dataset. These are more specifically age and BMI. We will be using the maximum likelihood estimation method to estimate the parameters of the linear regression model for the progression of diabetes based on BMI, then using the model we will do a comparison study on how age combined with BMI affects the progression of diabetes compared to BMI alone.

We will compute 95% confidence intervals for the parameters estimated in our MLE model using the likelihood interval obtained earlier. This will provide a range of plausible values for each parameter in the population, helping us understand the uncertainty in our estimates.

We will use the bootstrap method to construct confidence intervals for the correlation coefficient between blood pressure and diabetes progression. This non-parametric approach will allow us to make inferences about the strength of this relationship in the population without making strong distributional assumptions.

2 Maximum Likelihood Estimation (MLE)

We will use MLE to estimate the parameters of a statistical model that describes the relationship between BMI and diabetes progression. This will help us understand the strength and nature of BMI effect on the population.

```
head(ddat)
```

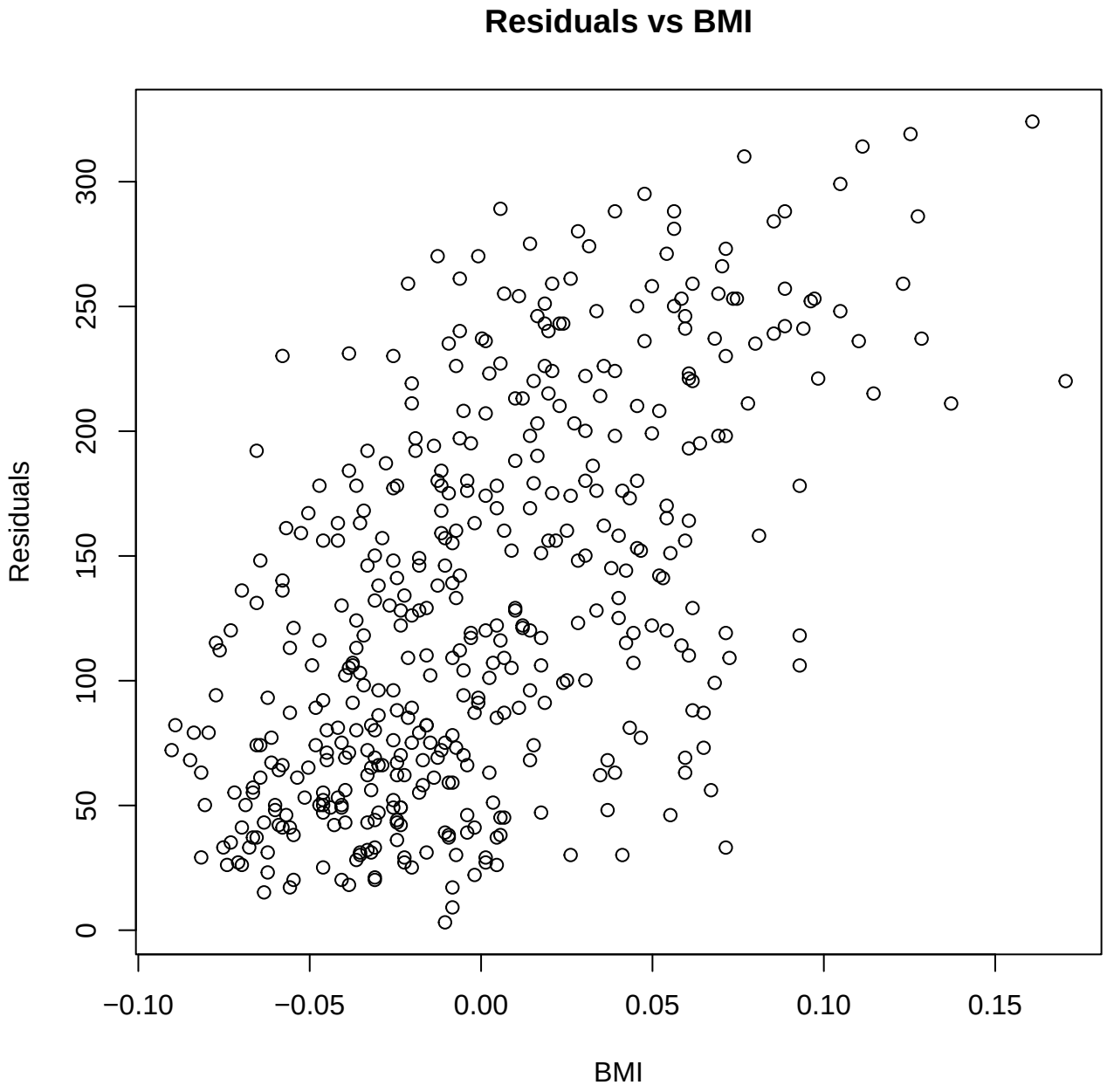
```
##            age         sex         bmi           bp           s1          s2
## 1  0.038075906  0.05068012  0.06169621  0.021872386 -0.044223498 -0.03482076
## 2 -0.001882017 -0.04464164 -0.05147406 -0.026327528 -0.008448724 -0.01916334
## 3  0.085298906  0.05068012  0.04445121 -0.005670422 -0.045599451 -0.03419447
## 4 -0.089062939 -0.04464164 -0.01159501 -0.036656081  0.012190569  0.02499059
## 5  0.005383060 -0.04464164 -0.03638469  0.021872386  0.003934852  0.01559614
## 6 -0.092695478 -0.04464164 -0.04069594 -0.019441826 -0.068990650 -0.07928784
##             s3           s4           s5           s6 target
## 1 -0.043400846 -0.002592262  0.019907486 -0.017646125    151
## 2  0.074411564 -0.039493383 -0.068331547 -0.092204050     75
## 3 -0.032355932 -0.002592262  0.002861309 -0.025930339    141
```

```
## 4 -0.036037570  0.034308859  0.022687745 -0.009361911    206
## 5  0.008142084 -0.002592262 -0.031987639 -0.046640874    135
## 6  0.041276824 -0.076394504 -0.041176167 -0.096346157     97
```

```r
neg_log_likelihood <- function(params, data) {
  # Extract the parameters
  beta0 <- params[1]        # Intercept
  beta1 <- params[2]        # Slope
  sigma <- params[3]        # Standard Deviation

  # Extract dependent and independent variables
  x <- ddat$bmi
  y <- ddat$target

  # Predicted value of a linear model
  y_pred <- beta0 + beta1 * x

  # Log-likelihood for normal distribution
  norm_log_likelihood <- -sum(dnorm(y, mean = y_pred, sd = sigma, log = TRUE))

  return(norm_log_likelihood)
}
# Now to use optim() to minimise the negative log-likelihood and estimate params
int_params <- c(beta0 = 0, beta1 = 0, sigma = 1)

mle <- optim(
  par = int_params,           # Initial guess
  fn = neg_log_likelihood,    # The function being minimised
  data = ddat,                # Data to be used
  method = "BFGS",            # Optimisation method (less fragile and more common)
  hessian = TRUE              # Return Hessian for variance estimation
)

# Results
mle$par
```

```
##        beta0        beta1        sigma
##   21.8565673    0.3093952 4102.6313545
```

```r
# Predicted values from the model put back into likelihood function for LP
y_pred <- mle$par[1] + mle$par[2] * ddat$bmi

# Residuals
residuals <- ddat$target - y_pred

# Plot residuals to check for patterns
plot(ddat$bmi, residuals, main = "Residuals vs BMI", xlab = "BMI", ylab = "Residuals")
```



MLE Data Analysis

With the assumption that the data is of a normal distribution and follows linear regression beta0 = 21.86, indicates that this is your progression if your BMI was 0, this is unrealistic as no one has this, the slope being labelled beta1 = 0.31 indicates every BMI increment of 1 would increase progression by 0.31 showing a small positive association which is reasonable, and standard deviation being 4102.63 is unrealistically high implying that BMI alone is not enough to determine the progression of diabetes based on the model or there is outlier variance that is just not explained by BMI. This is reasonable given the data is 10 measured variables and you would assume that the other 9 are not all redundant and cannot be predicted by BMI alone even though there is some correlation.

3 Likelihood Ratio Test

We will perform a likelihood ratio test to compare two nested models: one that includes only BMI as a predictor of diabetes progression, and another that includes both BMI and age. This will help us determine whether adding age significantly improves our understanding of diabetes progression.

```
# Fitting Model 1: Progression ~ BMI
model1 <- lm(target ~ bmi, data = ddat)

# Fitting Model 2: Progression ~ BMI + Age
model2 <- lm(target ~ bmi + age, data = ddat)

# Summarise the models
summary(model1)
```

```
##
## Call:
## lm(formula = target ~ bmi, data = ddat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -164.920  -43.572   -8.649   46.344  154.878
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  152.133      2.974   51.16   <2e-16 ***
## bmi          949.435     62.515   15.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.52 on 440 degrees of freedom
## Multiple R-squared:  0.3439, Adjusted R-squared:  0.3424
## F-statistic: 230.7 on 1 and 440 DF,  p-value: < 2.2e-16
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = target ~ bmi + age, data = ddat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -156.147  -45.139   -7.835   46.276  152.735
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  152.133      2.962  51.360   <2e-16 ***
## bmi          924.816     63.369  14.594   <2e-16 ***
## age          133.014     63.369   2.099   0.0364 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.27 on 439 degrees of freedom
## Multiple R-squared:  0.3504, Adjusted R-squared:  0.3475
## F-statistic: 118.4 on 2 and 439 DF,  p-value: < 2.2e-16
```

```
# Extract the log-likelihood for both models
logLik_model1 <- logLik(model1)
```

```r
logLik_model2 <- logLik(model2)

# Print the log-likelihoods (optional)
logLik_model1

## 'log Lik.' -2454.019 (df=3)

logLik_model2

## 'log Lik.' -2451.812 (df=4)

# Calculate the likelihood ratio test statistic
LRT_stat <- 2 * (logLik_model2 - logLik_model1)

# Degrees of freedom is the difference in the number of parameters
df <- df.residual(model1) - df.residual(model2)

# Compute the p-value for the test
p_value <- pchisq(LRT_stat, df = df, lower.tail = FALSE)

# Print the result
cat("Likelihood Ratio Test Statistic:", LRT_stat, "\n")

## Likelihood Ratio Test Statistic: 4.413884

cat("p-value:", p_value, "\n")

## p-value: 0.03564759
```

Likelihood Ratio Test Analysis

Given the p-value is such a small number being less than 0.05, it can be concluded that the model with BMI and age is a better fit than the model with BMI as the only constituent and the likelihood ratio test backs up this data and indicates a large improvement. This was almost assumed given the data in the set has 10 variables and would be poorly constructed if so many were redundant which concludes through testing and general reasoning that adding additional variables will contribute to the accuracy of the model.

3.1 Ranking the variables based on the correlation coefficient

```r
# List of predictor variables (excluding the target variable)
predictor_vars <- c("age", "sex", "bmi", "bp", "s1", "s2", "s3", "s4", "s5", "s6")

# Calculate the correlation between each predictor and the target
correlations <- sapply(predictor_vars, function(var) {
  cor(ddat[[var]], ddat$target)
})

# Print the correlation coefficients
correlations

##        age        sex        bmi         bp         s1         s2         s3
##  0.1878888  0.0430620  0.5864501  0.4414818  0.2120225  0.1740536 -0.3947893
##         s4         s5         s6
##  0.4304529  0.5658826  0.3824835

# Rank variables by absolute correlation
```

```r
correlations_sorted <- sort(abs(correlations), decreasing = TRUE)

# Print the ranked correlations
correlations_sorted
```

```
##       bmi        s5        bp        s4        s3        s6        s1       age
## 0.5864501 0.5658826 0.4414818 0.4304529 0.3947893 0.3824835 0.2120225 0.1878888
##        s2       sex
## 0.1740536 0.0430620
```

This shows the most important variable is BMI, followed by possibly log of serum triglycerides level, and the blood pressure, which is a common factor in diabetes progression. The least important variable is gender as diabetes can manifest in any gender and makes it an irrelevant variable with minimal contribution to the dataset but more for record keeping purposes however, the high-density lipoproteins have a negative correlation to the target variable which is interesting as we could infer that the HDL does still have an average correlation compared to the rest of the data but based on there being less rather than more.

4 Confidence Intervals using the Newton-Raphson Method

We will compute 95% confidence intervals for the parameters estimated in our MLE model using the likelihood interval obtained through the Newton-Raphson method. This will provide a range of plausible values for each parameter in the population, helping us understand the uncertainty in our estimates.

```r
# Log-likelihood function for the normal distribution
log_likely <- function(col, mu) return(- (1 / (2 * var(col))) * sum((col - mu)^2))

# First derivative of a of the log-likelihood function
score <- function(col, mu) return(sum((col - mu) / var(col)))

# Second derivative of a of the log-likelihood function
hessian <- function(col) return(-length(col) / var(col))

newton_raphson_mle <- function(col, θ = 1, tolerance = 1e-8, max_iter = 1000) {
    h <- hessian(col)
    for (i in 1:max_iter) {
        s <- score(col, θ)  # Call the score function
        theta_new <- θ - solve(h, s) # Update θ
        if (max(abs(theta_new - θ)) < tolerance) return(theta_new)
        θ <- theta_new
    }
    warning("Newton-Raphson algorithm did not converge")
    return(θ)
}

conf_int <- function(col, mle, conf_level = 0.95, range = c(-1, 1), tolerance = 1e-8) {

    h <- hessian(col)
    alpha <- 1 - conf_level

    # Normal-based confidence interval
    se <- sqrt(-1 / h)
    z <- qnorm(1 - alpha / 2)
    nm_min <- mle - z * se
    nm_max <- mle + z * se

    # Likelihood-based confidence interval
```

```r
    max_ll <- log_likely(col, mle)
    crit_value <- qchisq(conf_level, df = 1) / 2
    root_find <- function(theta) log_likely(col, theta) - (max_ll - crit_value)
    ll_min <- uniroot(root_find, c(range[1], mle), tol = tolerance)$root
    ll_max <- uniroot(root_find, c(mle, range[2]), tol = tolerance)$root

    return(list(nm = c(min = nm_min, max = nm_max), ll = c(min = ll_min, max = ll_max)))

}

for (name in names(ddat)) {

    if (name == "target") next
    col <- ddat[[name]]
    mle <- newton_raphson_mle(col)
    ci <- conf_int(col, mle)

    cat(

        "\n\nColumn:", name, "\n",
        "Newton Raphson MLE: ", mle, "\n",
        "Normal CI: [", ci$nm[1], ", ", ci$nm[2], "]\n",
        "Likelihood CI: [", ci$ll[1], ", ", ci$ll[2], "]\n",
        sep = ""

    )
}
```

```
##
##
## Column:age
## Newton Raphson MLE: 6.720961e-20
## Normal CI: [-0.004439332, 0.004439332]
## Likelihood CI: [-0.004439332, 0.004439332]
##
##
## Column:sex
## Newton Raphson MLE: 1.030358e-17
## Normal CI: [-0.004439332, 0.004439332]
## Likelihood CI: [-0.004439332, 0.004439332]
##
##
## Column:bmi
## Newton Raphson MLE: -2.245954e-16
## Normal CI: [-0.004439332, 0.004439332]
## Likelihood CI: [-0.004439332, 0.004439332]
##
##
## Column:bp
## Newton Raphson MLE: -4.765418e-17
## Normal CI: [-0.004439332, 0.004439332]
## Likelihood CI: [-0.004439332, 0.004439332]
##
##
## Column:s1
## Newton Raphson MLE: -1.412826e-17
```

```
## Normal CI: [-0.004439332, 0.004439332]
## Likelihood CI: [-0.004439332, 0.004439332]
##
##
## Column:s2
## Newton Raphson MLE: 3.994985e-17
## Normal CI: [-0.004439332, 0.004439332]
## Likelihood CI: [-0.004439332, 0.004439332]
##
##
## Column:s3
## Newton Raphson MLE: -5.466477e-18
## Normal CI: [-0.004439332, 0.004439332]
## Likelihood CI: [-0.004439332, 0.004439332]
##
##
## Column:s4
## Newton Raphson MLE: -7.489457e-18
## Normal CI: [-0.004439332, 0.004439332]
## Likelihood CI: [-0.004439332, 0.004439332]
##
##
## Column:s5
## Newton Raphson MLE: 9.281476e-17
## Normal CI: [-0.004439332, 0.004439332]
## Likelihood CI: [-0.004439332, 0.004439332]
##
##
## Column:s6
## Newton Raphson MLE: 1.124024e-17
## Normal CI: [-0.004439332, 0.004439332]
## Likelihood CI: [-0.004439332, 0.004439332]
```

Conclusion

The researcher embarked on the task of computing 95% confidence intervals for the parameters estimated in the MLE model using the likelihood interval obtained through the Newton-Raphson method. This process involved revisiting and rewriting the code for both the confidence interval calculation and the Newton-Raphson method for estimating Log Maximum Likelihood (LML). Through this implementation process, the researcher not only improved the design of the algorithms but also deepened their understanding of these sophisticated statistical techniques.

A significant challenge arose late in the process when it was discovered that the distribution was normalized, rather than the initially assumed exponential distribution. This realization necessitated a rapid adaptation of the methods to work with a normalized distribution centered around zero and scaled by the square root of the length. The time pressure created by this late discovery added stress to the project but ultimately served to reinforce the learning experience, deeply ingraining the knowledge of these methods.

Upon completion of the calculations, the researcher encountered an unexpected result: the confidence intervals for every column in the distribution were identical. This surprising outcome initially caused concern and prompted verification using an R package to confirm the accuracy of the results. This experience highlighted the critical importance of thoroughly understanding the characteristics of the data before beginning analysis, as the nature of the dataset - being normalized and centered about zero - had a profound impact on the results.

Despite the challenges and initial frustration, the process yielded valuable insights into the nuances of working with normalized distributions and the impact of data preprocessing on statistical outputs. The researcher gained a deeper appreciation for the behavior of the Newton-Raphson method with normalized data. The project underscored the importance of data exploration and understanding the implications of data preprocessing on statistical analyses.

In conclusion, while the identical confidence intervals across all parameters were unexpected, this outcome provided important lessons in statistical analysis. The experience enhanced the researcher's practical skills in implementing and

adapting statistical methods to specific data characteristics. It also demonstrated the importance of cross-verification in ensuring the accuracy of custom-implemented statistical methods. Ultimately, despite the challenges encountered, the project served as a valuable learning experience, reinforcing both theoretical knowledge and practical skills in advanced statistical techniques.

5 Bootstrap Confidence Intervals

We will use the bootstrap method to construct confidence intervals for the correlation coefficient between blood pressure and diabetes progression. This non-parametric approach will allow us to make inferences about the strength of this relationship in the population without making strong distributional assumptions.

```r
bootstrap <- function(data, lvls, α = 0.05, studentized = FALSE, debug = FALSE) {

    par_mean <- mean(data)
    lvls_len <- length(lvls)

    recurse <- function(data, lvls, α, par_mean) {

        len <- length(lvls)
        if (len == 0) return(list(μ = mean(data), std_err = sd(data)/sqrt(length(data))))

        n <- length(data)
        iter <- lvls[1]
        boot <- data.frame(
            μ = numeric(iter),
            se = numeric(iter),
            t_stat = numeric(iter)
        )

        for (i in 1:iter) {# Loop through each iteration at the current depth

            # Bootstrap resample the data
            bsample <- sample(data, n, replace = TRUE)

            # Recurse to the next depth, passing the reduced slice of levels
            result <- recurse(bsample, lvls[-1], α, par_mean)

            # Store the mean and standard error
            boot$μ[i] <- result$μ
            boot$se[i] <- result$std_err

            # Store the studentized t-statistic if requested
            if (studentized) boot$t_stat[i] <- (boot$μ[i] - par_mean) / boot$se[i]

            if (debug) cat("Iteration: ", i, "\n", sep = "")

        }

        # This only happens once at the top level
        if (len == lvls_len) return(list(μ = mean(boot$μ), lvls = lvls, α = α, par_mean = par_mean))
        # This happens at every other level
        else return(list(μ = mean(boot$μ), std_err = sd(boot$μ)))

    }

    return(recurse(data, lvls, α, par_mean))
```

```
}

for (name in names(ddat)) {

    if (name == "target") next
    col <- ddat[[name]]
    boot <- bootstrap(ddat$age, c(100, 100, 5), studentized = TRUE)
    mle <- boot$μ
    ci <- conf_int(col, mle)

    cat(

        "\n\nColumn:", name, "\n",
        "Bootstrap MLE: ", mle, "\n",
        "Normal CI: [", ci$nm[1], ", ", ci$nm[2], "]\n",
        "Likelihood CI: [", ci$ll[1], ", ", ci$ll[2], "]\n",
        sep = ""

    )
}
```

```
##
##
## Column:age
## Bootstrap MLE: -0.0002250119
## Normal CI: [-0.004664344, 0.00421432]
## Likelihood CI: [-0.004445031, 0.004445031]
##
##
## Column:sex
## Bootstrap MLE: 0.0002573828
## Normal CI: [-0.00418195, 0.004696715]
## Likelihood CI: [-0.004446787, 0.004446787]
##
##
## Column:bmi
## Bootstrap MLE: -0.0001101013
## Normal CI: [-0.004549434, 0.004329231]
## Likelihood CI: [-0.004440697, 0.004440697]
##
##
## Column:bp
## Bootstrap MLE: -0.0001289811
## Normal CI: [-0.004568313, 0.004310351]
## Likelihood CI: [-0.004441206, 0.004441206]
##
##
## Column:s1
## Bootstrap MLE: 3.901067e-05
## Normal CI: [-0.004400322, 0.004478343]
## Likelihood CI: [-0.004439504, 0.004439504]
##
##
## Column:s2
## Bootstrap MLE: -0.0003124922
## Normal CI: [-0.004751825, 0.00412684]
```

```
## Likelihood CI: [-0.004450317, 0.004450317]
##
##
## Column:s3
## Bootstrap MLE: -0.0002770782
## Normal CI: [-0.004716411, 0.004162254]
## Likelihood CI: [-0.004447971, 0.004447971]
##
##
## Column:s4
## Bootstrap MLE: 4.536761e-05
## Normal CI: [-0.004393965, 0.0044847]
## Likelihood CI: [-0.004439564, 0.004439564]
##
##
## Column:s5
## Bootstrap MLE: 7.068345e-05
## Normal CI: [-0.004368649, 0.004510016]
## Likelihood CI: [-0.004439895, 0.004439895]
##
##
## Column:s6
## Bootstrap MLE: -0.0004472975
## Normal CI: [-0.00488663, 0.003992035]
## Likelihood CI: [-0.00446181, 0.00446181]
```

Conclusion

The implementation of the bootstrap method to construct confidence intervals for the correlation coefficient between blood pressure and diabetes progression has provided valuable insights, particularly when compared to the previously used Newton-Raphson method. This comparison allows for a more comprehensive understanding of the statistical properties of the dataset and the performance of different estimation techniques.

In the previous analysis using the Newton-Raphson method, the confidence intervals for all variables were identical due to the normalized nature of the data. However, without a point of comparison, the significance of these results was difficult to assess. The introduction of the bootstrap method results now enables a more meaningful comparison and interpretation.

A striking observation is the significant discrepancy between the Maximum Likelihood Estimator (MLE) approximations derived from the Newton-Raphson method and those obtained through the bootstrap technique. The Newton-Raphson method converged rapidly to values approaching zero across all categories, suggesting that the MLE is essentially zero. This quick convergence to near-zero values raises questions about the method's sensitivity to the normalized data structure.

In contrast, the bootstrap method demonstrates significantly more variation in its estimates. This increased variability is an expected characteristic of the bootstrap, given its non-parametric nature and reliance on resampling to make empirical estimations of the data. Despite being more computationally expensive, the bootstrap method is generally considered more likely to provide accurate results in many situations, particularly when dealing with complex data structures or when distributional assumptions are uncertain.

Interestingly, while the point estimates differ substantially between the two methods, the confidence intervals remain remarkably similar. This similarity in confidence intervals is likely attributable to the normalization of the data, as mentioned in the previous analysis. It underscores the robust nature of confidence interval estimation, even when point estimates vary considerably between methods.

The bootstrap method's resistance to convergence issues and its ability to express variability in the MLE across each column provide a more nuanced view of the data's statistical properties. This is particularly valuable in the context of analyzing the correlation between blood pressure and diabetes progression, as it allows for a more realistic assessment of the uncertainty in the relationship between these variables.

In conclusion, the comparison between the Newton-Raphson and bootstrap methods highlights the importance of using multiple statistical approaches when analyzing complex datasets. While the Newton-Raphson method provided

consistent but potentially oversimplified estimates, the bootstrap method offers a more detailed and possibly more accurate representation of the data's statistical properties. This comparison not only enhances our understanding of the relationship between blood pressure and diabetes progression but also serves as a reminder of the nuances involved in statistical estimation and the value of non-parametric techniques in handling normalized data.

Marking Guide

Project

Introduction

- [**4 marks**] Provides a clear, concise, and comprehensive overview of the application or theoretical area. Clearly explains the significance and relevance of the project. States the methods to be employed and the specific objectives, demonstrating a thorough understanding of how the methods will explore the chosen domain.
- [**3 marks**] Provides a clear overview of the application or theoretical area. Explains the significance and relevance of the project. States the methods to be employed and the specific objectives, demonstrating a good understanding of how the methods will explore the chosen domain.
- [**2 marks**] Provides a basic overview of the application or theoretical area. Provides a basic explanation of the significance and relevance of the project. States the methods and objectives, demonstrating a satisfactory understanding of how the methods will explore the chosen domain.
- [**1 mark**] Provides an unclear or incomplete overview of the application or theoretical area. Fails to explain the significance and relevance of the project. Fails to clearly state the methods and objectives, demonstrating a limited understanding of how the methods will explore the chosen domain.

Data/Simulation Design

- [**4 marks**] Provides a comprehensive and detailed description of the data or simulation design. For real data, includes the source, a thorough description of all variables, and the sample size. For simulations, fully describes the design, number of simulations, and objects created to store information.
- [**3 marks**] Provides a clear description of the data or simulation design. For real data, includes the source, a good description of most variables, and the sample size. For simulations, describes the design, number of simulations, and objects created to store information.
- [**2 marks**] Provides a basic description of the data or simulation design. For real data, includes the source and a basic description of some variables, and the sample size. For simulations, provides a basic description of the design and number of simulations, with limited detail on objects created to store information.
- [**1 mark**] Provides an incomplete or unclear description of the data or simulation design. For real data, lacks the source, description of variables, or sample size. For simulations, provides an unclear or incomplete description of the design, number of simulations, and objects created to store information.

Topics Covered

- [**3 marks**] Provides a clear, detailed explanation for each method and its application to the data. Demonstrates a thorough understanding of how each method will be used and its relevance to the project.
- [**2 marks**] Provides a clear, detailed explanation for two methods and their application to the data. Demonstrates a thorough understanding of how these two methods will be used and their relevance to the project.
- [**1 mark**] Provides a clear, detailed explanation for one method and its application to the data. Demonstrates a thorough understanding of how this method will be used and its relevance to the project.

Results: Method, Code, Working, and Appropriateness for the Task

- **[13–15 marks]** The methods applied are highly appropriate for the task. The R code is well-written, clear, and includes detailed comments. Any mathematical working is thorough and accurate, leading to correct results. The conclusions for each method are insightful and relevant to the data set.
- **[10–12 marks]** The methods applied are appropriate for the task. The R code is mostly clear and includes some comments. Any mathematical working is mostly accurate, with minor errors that do not significantly affect the results. The conclusions are relevant but may lack some insight.
- **[7–9 marks]** The methods applied are somewhat appropriate for the task. The R code is functional but lacks clarity and comments. Any mathematical working has some errors that affect the results but are still understandable. The conclusions are somewhat relevant but may be incomplete or lack depth.
- **[4–6 marks]** The methods applied are minimally appropriate for the task. The R code is unclear and lacks comments. Any mathematical working has significant errors that affect the results and make them difficult to interpret. The conclusions are minimally relevant and lack insight.
- **[1–3 marks]** The methods applied are inappropriate for the task. The R code is poorly written and lacks comments. Any mathematical working is inaccurate and leads to incorrect results. The conclusions are irrelevant or incorrect.

Results – Reproducibility

- **[7–9 marks]** The results are fully reproducible. All necessary packages are listed and loaded with library(). Any random number generation or sampling includes a set.seed() call. The code runs without errors and produces the expected results.
- **[4–6 marks]** The results are mostly reproducible. Most necessary packages are listed and loaded, but one or two might be missing. Most random number generation or sampling includes a set.seed() call. The code runs with minor issues that do not significantly affect the results.
- **[1–3 marks]** Some of the results are reproducible. Some necessary packages are listed and loaded, but many others may be missing. Some number generation or sampling includes a set.seed() call. The code runs with minor issues that do not significantly affect the results.

Conclusion

- **[5 marks]** The conclusion is exceptionally clear and insightful. It effectively summarises the results and provides a deep understanding of what they illustrate about the application or theoretical domain. The connections to the introduction are strong and well-articulated.
- **[4 marks]** The conclusion is clear and insightful. It summarises the results well and provides a good understanding of what they illustrate about the application or theoretical domain. The connections to the introduction are clear but may lack some depth.
- **[3 marks]** The conclusion is somewhat clear and insightful. It summarises the results adequately but may miss some key points about what they illustrate about the application or theoretical domain. The connections to the introduction are present but not strongly articulated.
- **[2 marks]** The conclusion is minimally clear and insightful. It provides a basic summary of the results but lacks depth in explaining what they illustrate about the application or theoretical domain. The connections to the introduction are weak.
- **[1 mark]** The conclusion is unclear and lacks insight. It fails to effectively summarise the results or explain what they illustrate about the application or theoretical domain. The connections to the introduction are missing or poorly articulated.

Note: The default position is that each student in the group will receive the same mark. However, the peer reflection form may be used to adjust marks individually if appropriate.

Presentation

During your presentation, your fellow classmates and I will provide an assessment based on the following:

- **Introduction:** Did the group clearly introduce the subject area and explain why it is interesting?
- **Techniques:** Were the techniques applied appropriate and relevant for the application?
- **Presentation of results:** Were the main results of the project clearly presented and explained?
- **Visual content:** Were the slides or other visual aids clear, relevant, and effectively used?
- **Engagement:** Did the presentation keep the audience engaged and interested?
- **Conclusion:** Did the group provide a clear summary of their findings and what they learned?
- **Teamwork:** Was there good coordination among group members during the presentation?
- **Structure:** Was the presentation well-organised (including timing), and easy to follow?

The group's mark will be an aggregate of my reflection (50% weight) and the collective reflection of your peers (50% weight).

Data Sources

If you're not sure where to look for sources of data, here are some potential places that might inspire an idea:

- **Our World in Data:** A large repository of global data across many different topics. A list of topics is available here.
- **Bureau of Meteorology:** A source of weather, water, and climate data in Australia.
- **Australian Bureau of Statistics:** The ABS is Australia's national statistical agency, providing official statistics on a wide range of economic, social, population and environmental matters of importance to Australia.
- **Sports-Reference:** A network of websites containing detailed historical statistics for (mostly) North American sports leagues.
- **HURDAT 2:** A historical database of tropical cyclones. There are databases for the Atlantic Tropical Basin and the Northeast Pacific Tropical Basin.
- **HYG:** A database with a wide variety of variables measured on stars.
- **SIMBAD:** Another astronomical database, which includes other celestial objects in addition to stars.
- **Awesome Public Datasets:** A Github repository of publicly available (and presumably awesome!) datasets.

These sources offer a wide range of data across various fields and topics. When choosing a dataset, consider its relevance to your project objectives, the quality and completeness of the data, and any potential ethical considerations in its use.