1. Review

We have discussed simple linear regression. The goal of regression is to describe the behavior of some response variable by the behavior of predictor (or explanatory) variables. For simple linear regression, we have n independent subjects with a single predictor. We will extend these methods to incorporate more predictor variables later. Here are some review questions about simple linear regression.

a. What is the model that we fit?

The model that we fit is the simple linear regression model $Y_i = β_0 + β_1X_i + ϵ_i$; where Y_i is the value of the response variable in the i-th observation. X_i is the value of the predictor variable in the i-th observation. $β_0$ and $β_1$ are parameters, and $ϵ_i$ is a random 'error' term

Estimation

b. What are two methods for fitting this model?

Two methods for fitting the simple linear regression model are the method of least squares estimation and the method of maximum likelihood. The method of least squares estimation considers the sum of the n squared deviations, denoted by Q, to find estimations of the regression parameters $β_0$ and $β_1$. The method of maximum likelihood selects values of the model parameters that are most consistent with the sample data using the likelihood function L(μ).

c. What assumptions do these methods make?

The method of least squares has three major assumptions. The first is that the error term $ϵ_i$ has a conditional mean of zero given X_i: $E(ϵ_i \mid X_i) = 0$. The second assumption is that (X_i, Y_i), i = 1, … , n are independent and identically distributed draws from their joint distribution. Finally, large outliers are unlikely. X_i and Y_i have nonzero finite fourth moments (Arnold & Hanck, 2020). The method of maximum likelihood has two assumptions. The first assumption is that the data must be independently distributed, and the second assumption is that the data must be identically distributed (Eppes, 2019).

d. How do they differ and how are they the same?

There are several similarities between the two methods. They rely on similar underlying assumptions, including that the data is independently and identically distributed. Additionally, both methods rely on the slope and intercept parameters. Finally, both models aim to minimize differences between the observed values and the predicted values to create an optimized model.

There are also differences between the two methods. Least squares attempts to minimize the sum of the squared residuals, whereas the maximum likelihood method tries to maximize the likelihood function. The maximum likelihood function is considered more efficient with smaller variance when compared with the least squares method. Finally, each method is more appropriate in different contexts, depending on the data characteristics and the most appropriate assumptions.

Inference

e. What test do we use to test if the coefficient $β_1$ (the regression coefficient of the predictor) is different from 0? Why would we want to test if $β_1$ is equal to 0?

A t-test is used to test if the regression coefficient of the predictor is different from zero. It tests if the coefficient is different from zero to determine if there is a statistically significant relationship between the predictor variable and the dependent variable. We want to test if $β_1$ is equal to 0 to determine the importance of the predictor variable, test if the null hypothesis applies or should be rejected, and determine if the predictor variable should be included in the model or not.

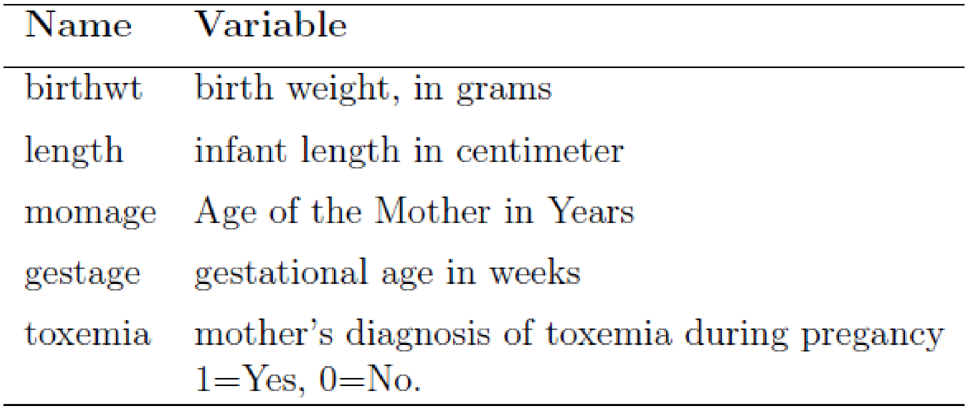
Confidence and Prediction Intervals

f. Confidence interval is for the true mean value of the response variable given a specific value of the predictor. The prediction interval is for the individual response variable value given a specific value of the predictor, and thus involves additional variability compared to the confidence interval and is much wider than the confidence interval.

No question present.

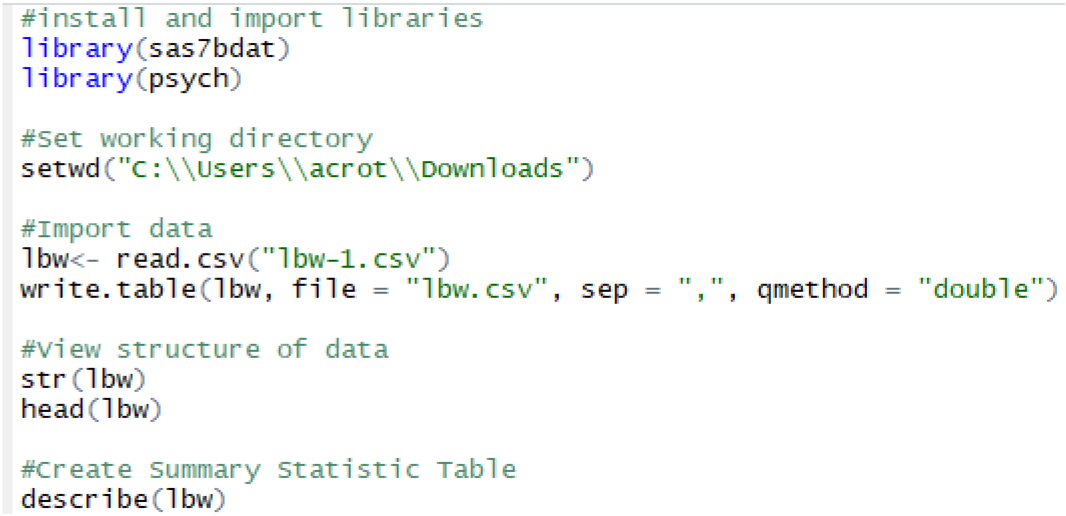
2. Example

We will use the low birth weight data set as the example. Recall that the data set contains information from a random sample of 100 low birth weight infants born in Boston, MA in 1990s. The response (outcome) variable of interest is headcirc, head circumference measurements in centimeters. Other variables in the dataset are described in the following table. The data set named lbw.sas7bdat is posted on the course web page in Moodle.

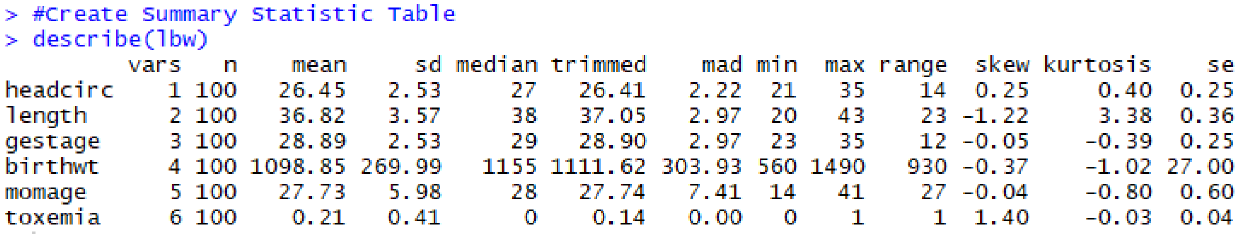


Name	Variable
birthwt	birth weight, in grams
length	infant length in centimeter
momage	Age of the Mother in Years
gestage	gestational age in weeks
toxemia	mother's diagnosis of toxemia during pregancy 1=Yes, 0=No.

a. First, Numerical summary of the data.



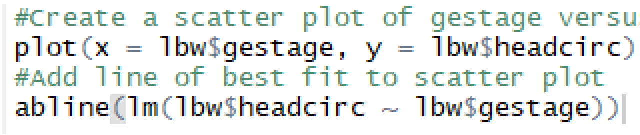
```
#install and import libraries
library(sas7bdat)
library(psych)

#Set working directory
setwd("C:\\Users\\acrot\\Downloads")

#Import data
lbw<- read.csv("lbw-1.csv")
write.table(lbw, file = "lbw.csv", sep = ",", qmethod = "double")

#View structure of data
str(lbw)
head(lbw)

#Create Summary Statistic Table
describe(lbw)
```



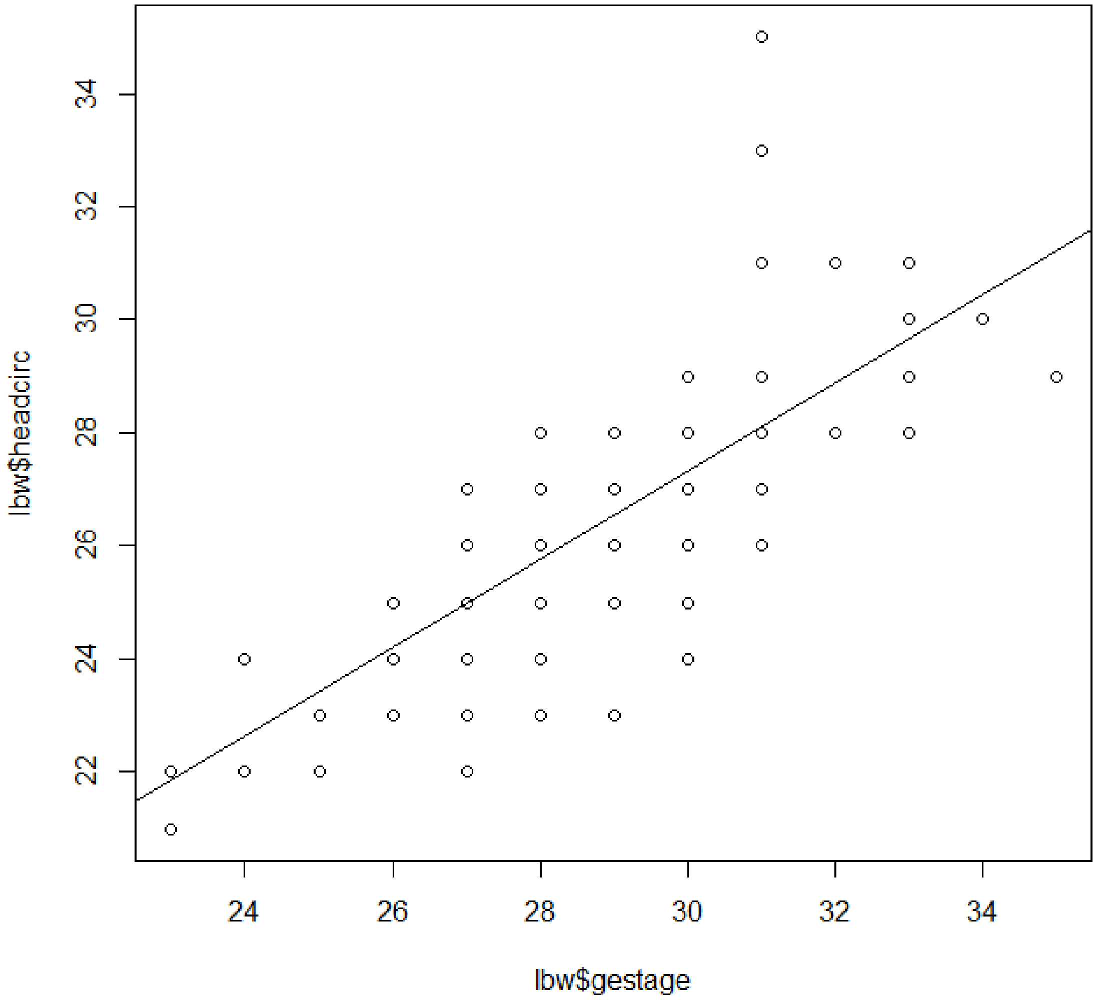
```
> #Create Summary Statistic Table
> describe(lbw)
          vars   n    mean     sd median trimmed    mad min  max range  skew kurtosis    se
headcirc     1 100   26.45   2.53     27   26.41   2.22  21   35    14  0.25     0.40  0.25
length       2 100   36.82   3.57     38   37.05   2.97  20   43    23 -1.22     3.38  0.36
gestage      3 100   28.89   2.53     29   28.90   2.97  23   35    12 -0.05    -0.39  0.25
birthwt      4 100 1098.85 269.99   1155 1111.62 303.93 560 1490   930 -0.37    -1.02 27.00
momage       5 100   27.73   5.98     28   27.74   7.41  14   41    27 -0.04    -0.80  0.60
toxemia      6 100    0.21   0.41      0    0.14   0.00   0    1     1  1.40    -0.03  0.04
```

b. Then, We will begin by analyzing the effect of gestage on headcirc.

i. Draw a scatter plot of gestage versus headcirc.



```
#Create a scatter plot of gestage versus headcirc
plot(x = lbw$gestage, y = lbw$headcirc)
#Add line of best fit to scatter plot
abline(lm(lbw$headcirc ~ lbw$gestage))
```



ii. What is the model we will fit?

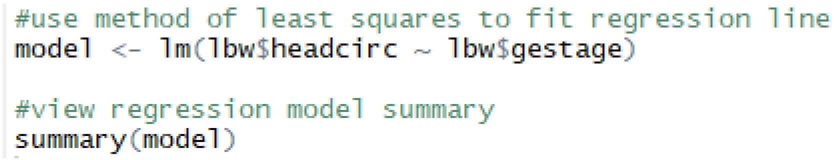
The model that we fit is the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$.

iii. What is the estimate for the effect of gestage on headcirc? How do you interpret this?

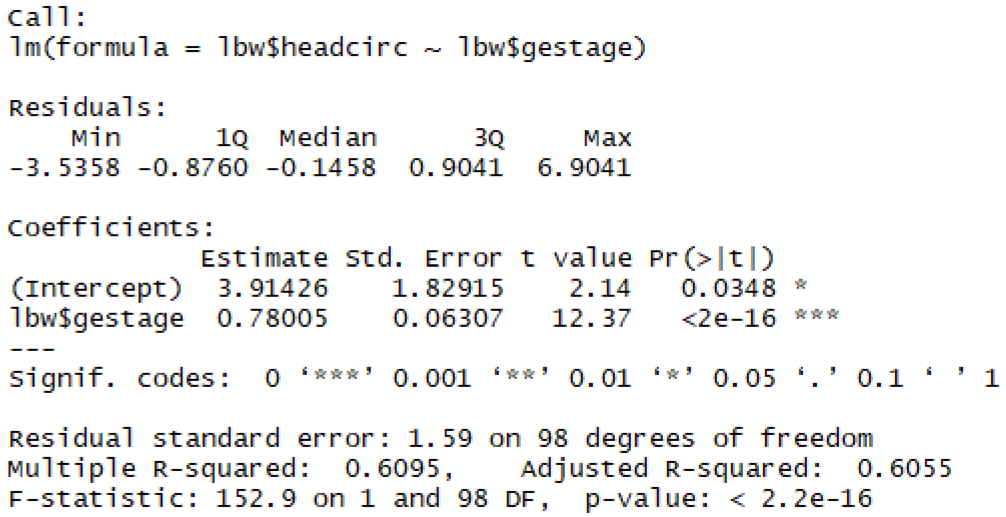
Based on the line of best fit, there appears to be an increase in the measured head circumference, in centimeters, based on the gestational age, in weeks. This makes sense, as while every fetus will grow at a different rate based on several factors, generally the longer the gestation, the farther along the fetus will be in terms of growth overall, including head circumference.

c. We will now proceed to make inferences based on the fitted model.

i. Perform the appropriate t-test to determine if there is a significant relationship between gestage and headcirc?



```
#use method of least squares to fit regression line
model <- lm(lbw$headcirc ~ lbw$gestage)

#view regression model summary
summary(model)
```



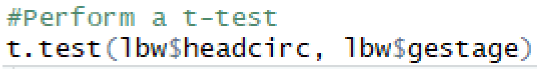
```
Call:
lm(formula = lbw$headcirc ~ lbw$gestage)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5358 -0.8760 -0.1458  0.9041  6.9041 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.91426    1.82915    2.14   0.0348 *  
lbw$gestage  0.78005    0.06307   12.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.59 on 98 degrees of freedom
Multiple R-squared:  0.6095,	Adjusted R-squared:  0.6055 
F-statistic: 152.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

Based on the regression model summary, the t-test statistic for the slope β_1 is 12.37 and its p-value is <2e-16. The size of the test statistic indicates a large difference between the observed relationship between the variables and the null hypothesis. Using the typical threshold of 0.05, we can conclude that there is significant evidence to reject the null hypothesis. Therefore, this test indicates a strong linear relationship between gestational age and head circumference.

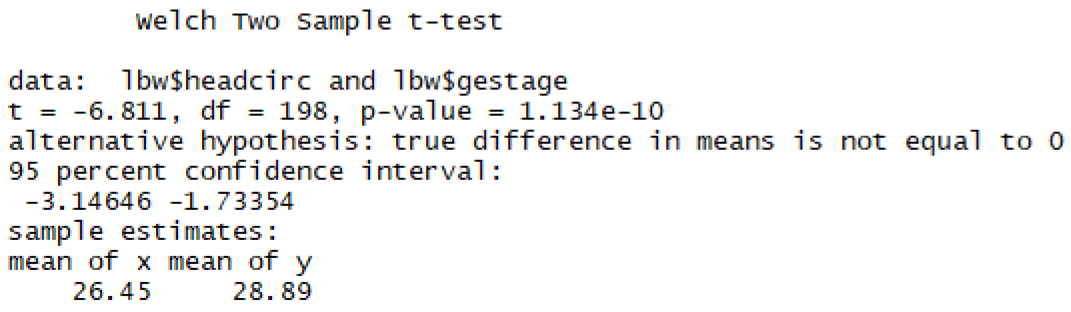
ii. How does this compare to the F-test result given in the output?

The F-test confirms and aligns with the results of the t-test in the output. The high F-test statistic of 152.9 indicates a strong fit of the regression model. The p-value of <2.2e-16 also falls below the 0.05 threshold, again indicating significant evidence to reject the null hypothesis.

iii. What if we were only interested in testing if increased gestage lead to an increase in headcirc? Perform this test.



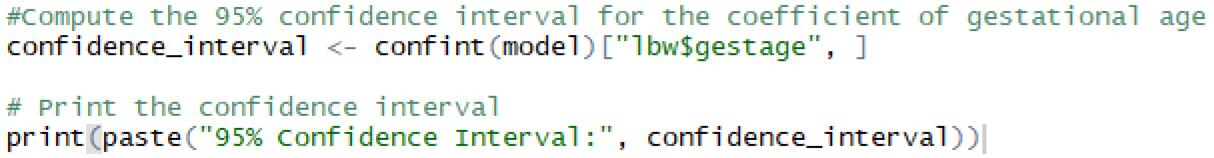
```
#Perform a t-test
t.test(lbw$headcirc, lbw$gestage)
```



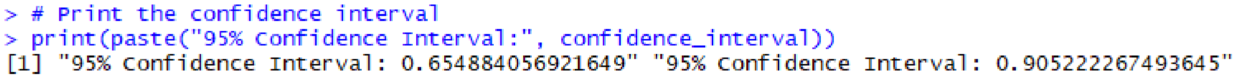
```
        Welch Two Sample t-test

data:  lbw$headcirc and lbw$gestage
t = -6.811, df = 198, p-value = 1.134e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.14646 -1.73354
sample estimates:
mean of x mean of y
    26.45     28.89
```

By definition, the simple linear regression tests this using the slope and regression coefficients. Performing a t-test can confirm these results. The results of this test support the alternative hypothesis that there is a significant difference in head circumference based on the duration of gestation, or gestational age. Therefore, an increase in gestational age indicates a statistically significant chance that the head circumference will increase.

iv. Find a two-sided 95% confidence interval for 1, the regression coefficient of gestage.



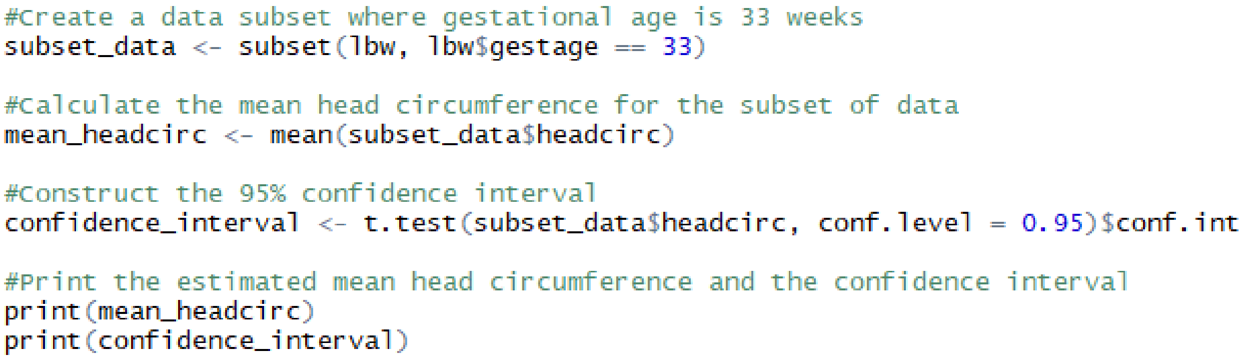
```
#Compute the 95% confidence interval for the coefficient of gestational age
confidence_interval <- confint(model)["lbw$gestage", ]

# Print the confidence interval
print(paste("95% Confidence Interval:", confidence_interval))
```



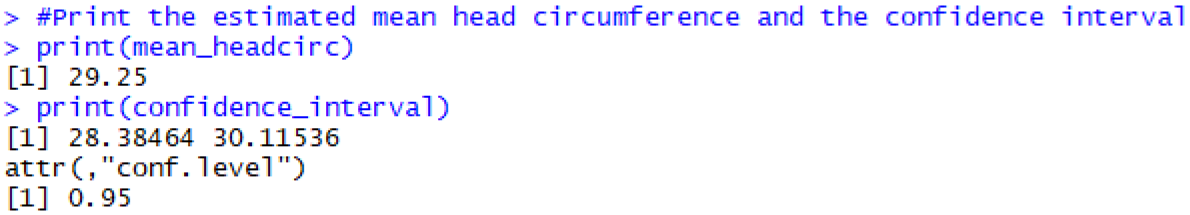
```
> # Print the confidence interval
> print(paste("95% Confidence Interval:", confidence_interval))
[1] "95% Confidence Interval: 0.654884056921649" "95% Confidence Interval: 0.905222267493645"
```

The 95% confidence interval is (0.6549, 0.9052). For a gestational age of 1 week, we can say with 95% confidence that the head circumference will be between 0.6549 cm and 0.9052 cm.

v. Find a two-sided 95% confidence interval of the mean value of headcirc for those with a gestage of 33 weeks.



```
#Create a data subset where gestational age is 33 weeks
subset_data <- subset(lbw, lbw$gestage == 33)

#Calculate the mean head circumference for the subset of data
mean_headcirc <- mean(subset_data$headcirc)

#Construct the 95% confidence interval
confidence_interval <- t.test(subset_data$headcirc, conf.level = 0.95)$conf.int

#Print the estimated mean head circumference and the confidence interval
print(mean_headcirc)
print(confidence_interval)
```



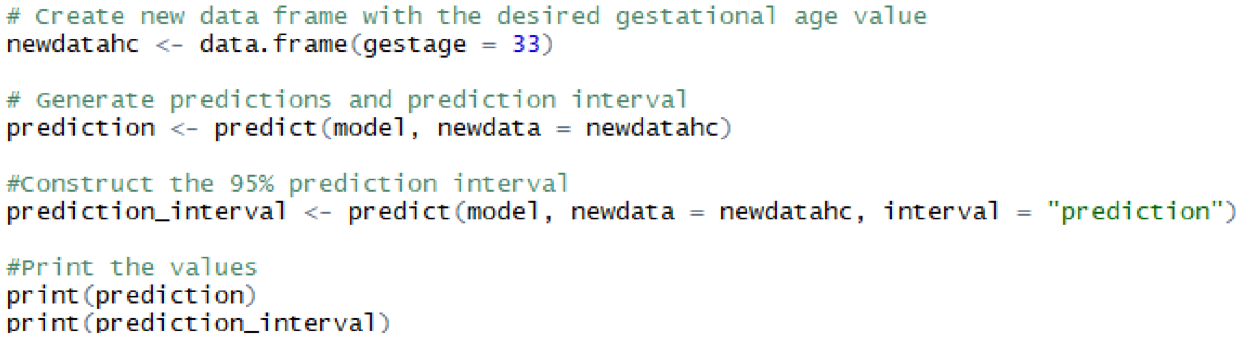
```
> #Print the estimated mean head circumference and the confidence interval
> print(mean_headcirc)
[1] 29.25
> print(confidence_interval)
[1] 28.38464 30.11536
attr(,"conf.level")
[1] 0.95
```

The estimated mean head circumference for a gestational age of 33 weeks is 29.25 cm. The 95% confidence interval for the true mean value of head circumference at 33 weeks is (28.38464, 30.11536).

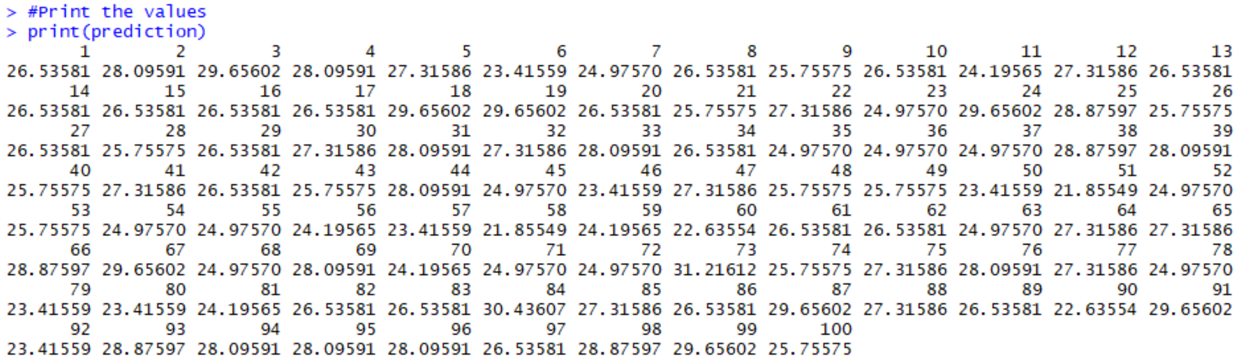
vi. How do you interpret this interval?

This interval can be interpreted as we are 95% confident that at 33 weeks of gestational age, the average head circumference will fall between 28.38464 cm and 30.11536 cm.

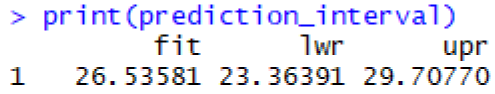
vii. Calculate the prediction interval of headcirc for a future observation with gestage of 33 weeks.



```
# Create new data frame with the desired gestational age value
newdatahc <- data.frame(gestage = 33)

# Generate predictions and prediction interval
prediction <- predict(model, newdata = newdatahc)

#Construct the 95% prediction interval
prediction_interval <- predict(model, newdata = newdatahc, interval = "prediction")

#Print the values
print(prediction)
print(prediction_interval)
```



```
> #Print the values
> print(prediction)
       1        2        3        4        5        6        7        8        9       10       11       12       13 
26.53581 28.09591 29.65602 28.09591 27.31586 23.41559 24.97570 26.53581 25.75575 26.53581 24.19565 27.31586 26.53581 
      14       15       16       17       18       19       20       21       22       23       24       25       26 
26.53581 26.53581 26.53581 26.53581 29.65602 29.65602 26.53581 25.75575 27.31586 24.97570 29.65602 28.87597 25.75575 
      27       28       29       30       31       32       33       34       35       36       37       38       39 
26.53581 25.75575 26.53581 27.31586 28.09591 27.31586 28.09591 26.53581 24.97570 24.97570 24.97570 28.87597 28.09591 
      40       41       42       43       44       45       46       47       48       49       50       51       52 
25.75575 27.31586 26.53581 25.75575 28.09591 24.97570 23.41559 27.31586 25.75575 25.75575 23.41559 21.85549 24.97570 
      53       54       55       56       57       58       59       60       61       62       63       64       65 
25.75575 24.97570 24.97570 24.19565 23.41559 21.85549 24.19565 22.63554 26.53581 26.53581 24.97570 27.31586 27.31586 
      66       67       68       69       70       71       72       73       74       75       76       77       78 
28.87597 29.65602 24.97570 28.09591 24.19565 24.97570 24.97570 31.21612 25.75575 27.31586 28.09591 27.31586 24.97570 
      79       80       81       82       83       84       85       86       87       88       89       90       91 
23.41559 23.41559 24.19565 26.53581 26.53581 30.43607 27.31586 26.53581 29.65602 27.31586 26.53581 22.63554 29.65602 
      92       93       94       95       96       97       98       99      100 
23.41559 28.87597 28.09591 28.09591 28.09591 26.53581 28.87597 29.65602 25.75575 
```



```
> print(prediction_interval)
       fit      lwr      upr
1 26.53581 23.36391 29.70770
```

The predicted head circumference for this child is 26.53581. The 95% confidence interval for this new value is (23.36391,29.70770).

viii. How do you interpret this prediction interval?

The predicted value falls within the confidence interval range. The confidence interval allows us to say that 95% of the infants with a gestational age of 33 weeks will have a head circumference that falls within 23.36391 cm and 29.70770 cm with variation in data taken into account. A true measurement of head circumference will reasonably fall within those values.

References

Arnold, M., & Hanck, C. (2020, September 15). *Introduction to econometrics with R*. 4.4 The Least Squares Assumptions. https://www.econometrics-with-r.org/4-4-tlsa.html

Eppes, M. (2019, September 21). *Maximum likelihood estimation explained - normal distribution*. Medium. https://towardsdatascience.com/maximum-likelihood-estimation-explained-normal-distribution-6207b322e47f