1. (55 points) The investigators are interested in assessing the relationship between Systolic Blood Pressure (SBP) in mmHg and Age in years among Hypertensive Patients. Specifically, whether a patient's SBP can be predicted from his or her age. They selected n=122 patients at random from a medical record database in a hospital. Assume that the simple linear regression model is appropriate.
The following table shows regression output of a simple linear regression model relating the SBP to the predictor Age. A few items have been masked and replaced by 'xxxx'.

Table 1: Least Squares Regression Results for the SBP Data

Analysis of Variance (ANOVA) Table

| Source of Variation | DF | Sum of Squares | Mean Square | $F$-Value | p-value |
|---|---|---|---|---|---|
| *Regression* | xxxx | xxxx | xxxx | xxxx | xxxx |
| *Error* | xxxx | xxxx | xxxx | | |
| *Total* | xxxx | xxxx | | | |

Parameter Estimates Table

| Variable | Parameter Estimates | Standard Error | $t$-Value | p-value |
|---|---|---|---|---|
| *Intercept* | 131.16 | 1.853 | xxxx | $< 0.0001$ |
| *Age* | xxxx | 0.028 | 15.27 | $< 0.0001$ |

sample size $n$=122, $R^2 = 0.6603$, $\hat{\sigma}^2$=29.8447

Answer the following questions based on the information provided in Table 1.

1. (5 points) Let Y be SBP and X be Age of a patient, write out the simple linear regression model. What are the assumptions of the model?

The simple linear regression model is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, 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. $\beta_0$ and $\beta_1$ are parameters, and $\epsilon_i$ is a random 'error' term. The assumptions of the model are that the regression line is straight, the expected value of the errors is zero, the variance of the errors is constant, and the errors are uncorrelated with each other, therefore, the outcome measures are uncorrelated with each other.

2. (4 points) What are the least squares estimates for intercept and slope, i.e., $\hat{\beta}_0$ and $\hat{\beta}_1$?

The intercept estimate $\hat{\beta}_0$: This estimated expected value of the response variable when the predictor equals zero. The intercept estimate (constant term) is 131.16. This is the estimated SBP when the age is 0 years if X = 0 is in the range of the data.
The slope estimate $\hat{\beta}_1$: This estimated expected change in the response variable associated with a one unit change in predictor. We use the formula:

$\beta_1$ = t-value * standard error
$\beta_1$ = 15.27 * 0.028 = 0.42756

3. (4 points) Give an interpretation of both $\beta\hat{}_0$ and $\beta\hat{}_1$.

Since $\beta\hat{}_0$ represents the SBP when the age is 0 years, it is probably not meaningful in the terms of this examination. However, it is still included for the data in order to create the estimated regression line. $\beta\hat{}_1$ represents the increase in systolic blood pressure for each increase in year of age, by 0.42756. Context is important, and other factors may also influence measurements, especially in pediatric patients.

4. (5 points) Write out the estimated regression line (function), and then calculate estimated expected value (or mean value) of SBP for a patient of age 70 years old.

Y = β^0 + β^1*X
SBP = 131.16 + 0.42756 * Age
SBP = 131.16 + 0.42756 * 70
SBP = 161.0892
The expected (mean) value of SBP for a patient of age 70 years old is 161.0892.

5. (3 points) What is the estimated expected change in SBP associated with a 5 year increase in Age?

Change in SBP = β^1 * Increase in Age
= 0.42756 * 5
= 2.1378
There is an estimated increase of 2.1378 associated with a 5-year increase in age.

6. (16 points) Complete the 8 missing numbers in 'DF', 'Sum of Squares' and 'Mean Square' in the ANOVA table above. You must show the intermediate steps of how you get results for these missing numbers.

**Regression DF:** The number of explanatory variables (including the intercept) in the regression model. This is 1.
Regression Sum of Squares: The sum of the squared differences between the predicted values and the mean of the response variable.
SSR = SST - SSE
Regression Mean Square: MSR = SSR/1, where 1 is its degrees of freedom.
**Error DF:** The total number of observations minus the DF for regression.
DF for error = n - 1 = 122 - 1 = 121
Error Sum of Squares: The sum of the squared differences between the observed values and the predicted values from the regression model.
SSE = (1 - R-squared) * SST
Error Mean Square: MS for error = SSE / DF for error
**Total DF**: the total number of observations minus 1. In a simple linear regression model, DF for total is given by n - 1, where n is the total number of observations. 122 - 1 = 121.
Total Sum of Squares: This represents the sum of the squared differences between the observed values and the mean of the response variable.
Sums of Squares Total (SST) = Sums of Squares Regression (SSR) + Sums of Squares Error (SSE)

7. (6 points) Calculate the F test statistic and its p-value in ANOVA table. State the null and alternative hypothesis here, the degrees of freedom of F test statistic, and make a conclusion at significance level $\alpha = 0.05$. (Using F-distribution table at Appendix B.4 of the textbook to find p-value)

F-test statistic = MSR / MSE
H0: $\beta 1 = 0$ (There is no linear relationship between Age and SBP)
Ha: $\beta 1 \neq 0$ (There is a linear relationship between Age and SBP)
If p-value < $\alpha$ (p-value < 0.05), we reject the null hypothesis.
If p-value >= $\alpha$ (p-value >= 0.05), we fail to reject the null hypothesis.

8. (3 points) Can you calculate the variance of response variable SBP (Y) using the available information? If so, provide it. If not, explain why.

9. (2 points) What is the relative reduction in the variation of Y when X is introduced into the regression model?

R-squared measures the total variation in the response variable of the model. In the provided table, R-squared is 0.6603, indicating 66.03% of the variation can be explained by age. This means age accounts for a significant amount of the variability in the regression model.

10. (4 points) Suppose we want to measure the association between SBP and Age using Pearson correlation coefficient. Can you calculate Pearson correlation coefficient between SBP and Age using the available information? If so, provide it. If not, explain why. Is it a positive or negative association? Why?

11. (3 points) Calculate the t test statistic for intercept $\beta_0$. What are the degrees of freedom for this t test?

t-test statistic = (β^0) / (SE)
t-test statistic = 131.16 / 1.853 = 70.78
DF = n - 2 = 122 - 2 = 120
The t-test statistic is 70.78, and the degrees of freedom is 120.

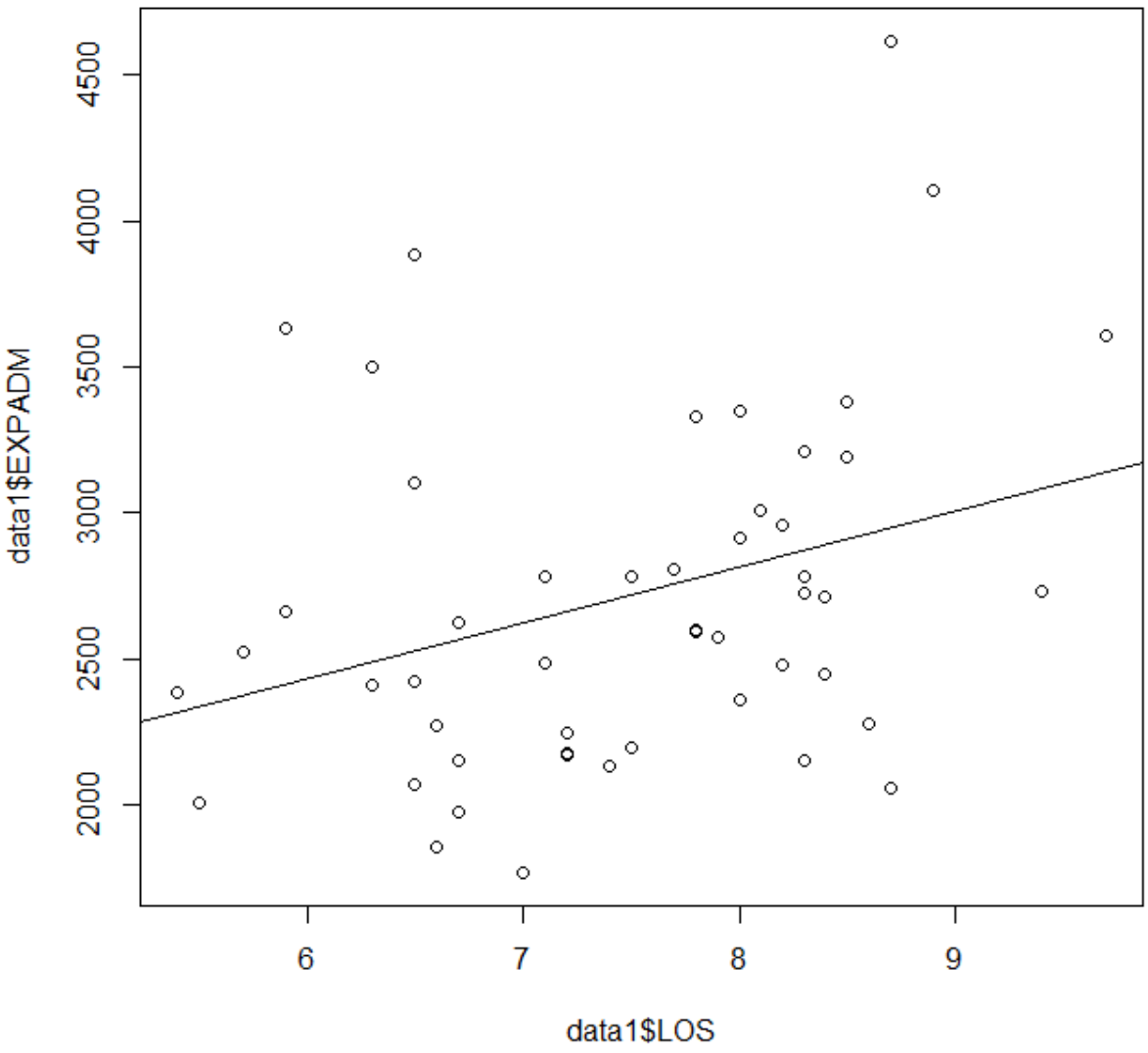
2. (22 points) Statistics that summarize personal health care expenditures by state for the years 1966 through 1982 have been examined in an attempt to understand issues related to rising health care costs. Suppose that you are interested in focusing on the relationship between expense per admission into a community hospital and average length of stay in the facility. The data set hospital.sas7bdat contains information for each state in the United States (including the District of Columbia) for the year 1982. The measures of mean expense per admission are saved under the variable name expadm; and the corresponding average lengths of stay are saved under los.
    a. (6 points) Using SAS or other statistical software you prefer, obtain numerical summary statistics (i.e., mean, median, range, standard deviation, etc.) for the variables expense per admission and length of stay in the hospital. Tabulate the summary statistics in a table, and then write a short paragraph describing the results.

```r
#install and import libraries
#install.packages("psych")
library(psych)

#import data
data1 <- read.csv('C:\\Users\\acrot\\Downloads\\hospital.csv')

#view structure of data
str(data1)
head(data1)

#Create summary statistic table for expadm and los variables
describe(data1[ , c('EXPADM', 'LOS')])
```

```
> #Create summary statistic table for expadm and los variables
> describe(data1[ , c('EXPADM', 'LOS')])
       vars  n    mean     sd median trimmed    mad    min    max  range  skew kurtosis    se
EXPADM    1 51 2716.80 603.95 2600.0 2659.54 598.97 1772.0 4612.0 2840.0  0.90     0.54 84.57
LOS       2 51    7.49   1.02    7.7    7.52   1.04    5.4    9.7    4.3 -0.16    -0.76  0.14
>
```

For the expense per admission variable, there are 51 observations. The mean expense per admission is $2,716.80, with a standard deviation of $603.95. The median is $2,600.00. The trimmed mean (removing 10% of observations from each end) is $2,659.54. The median absolute deviation is $598.97. The minimum visit was $1,772.00 and the maximum was $4,612.00. The range of values spans $2,840.00. There is a skewness value of 0.90 (slightly skewed) with kurtosis of 0.54 (more peaked than normal). The standard error is $84.57. For the length of stay variable, there are 51 observations. The mean length is 7.49 days, with a standard deviation of 1.02 days. The median is 7.7 days. The trimmed mean (removing 10% of observations from each end) is 7.52 days. The median absolute deviation is 1.04 days. The minimum stay was 5.4 days and the maximum was 9.7 days. The range of values spans 4.3 days. There is a skewness value of -0.16 (nearly symmetrical) with kurtosis of -0.76 (flatter shape than normal). The standard error is 0.14 days.

b. (4 points) Use appropriate graphic method to explore the relationship between expense per admission versus length of stay. Write one or two sentence on what you find about the nature of the relationship between these variables?




Based on the line of best fit, there appears to be an increase in the expense per admission based on a higher length of stay. This makes sense, as a higher length of stay would result in additional costs including medication, treatments, and staffing costs. However, there appears to be many scatter points far from the line of best fit, indicating a lot of variation in cost. This also makes sense, as some procedures and medications will invoke greater cost compared to others.

c. (4 points) Using expense per admission as the response and length of stay as the explanatory or predictor variable, compute the least-squares regression line using software. Also, interpret the estimated slope and intercept of the estimated line in the context of the problem.

```
#use method of least squares to fit regression line
model <- lm(data1$EXPADM ~ data1$LOS)

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

```
Call:
lm(formula = data1$EXPADM ~ data1$LOS)

Residuals:
   Min     1Q Median     3Q    Max
-889.6 -428.1 -102.1  265.8 1663.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1281.96     608.10   2.108   0.0402 *
data1$LOS     191.56      80.47   2.381   0.0212 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 577.6 on 49 degrees of freedom
Multiple R-squared:  0.1037,    Adjusted R-squared:  0.08538
F-statistic: 5.668 on 1 and 49 DF,  p-value: 0.02121
```

The least-squares regression line is 1281.96 + 191.56(length of stay). The estimated slope is 1281.96, and the intercept is 191.56. This means that for someone with a length of stay of 0 days, they can still expect a bill of $1,281.96. For each additional day of stay, they can expect the bill to increase by $191.56.

d. (3 points) What is the 95% confidence interval for the slope $\beta_1$, the slope of the population regression line. What does this interval tell you about the linear relationship between expense per admission and length of stay in the hospital?

```
> #Find 95% confidence interval for the slope
> confint(model,level=0.95)
                2.5 %    97.5 %
(Intercept) 59.92851 2503.9904
data1$LOS   29.86172  353.2643
```

The 95% confidence interval for the slope $\beta_1$ is (29.86172, 353.2643). This means that the linear relationship between expense per admission and length of stay in the hospital is statistically significant. For each additional day in the length of stay, we can say with 95% confidence that the expense per admission is expected to increase by an amount between $29.86 and $353.26.

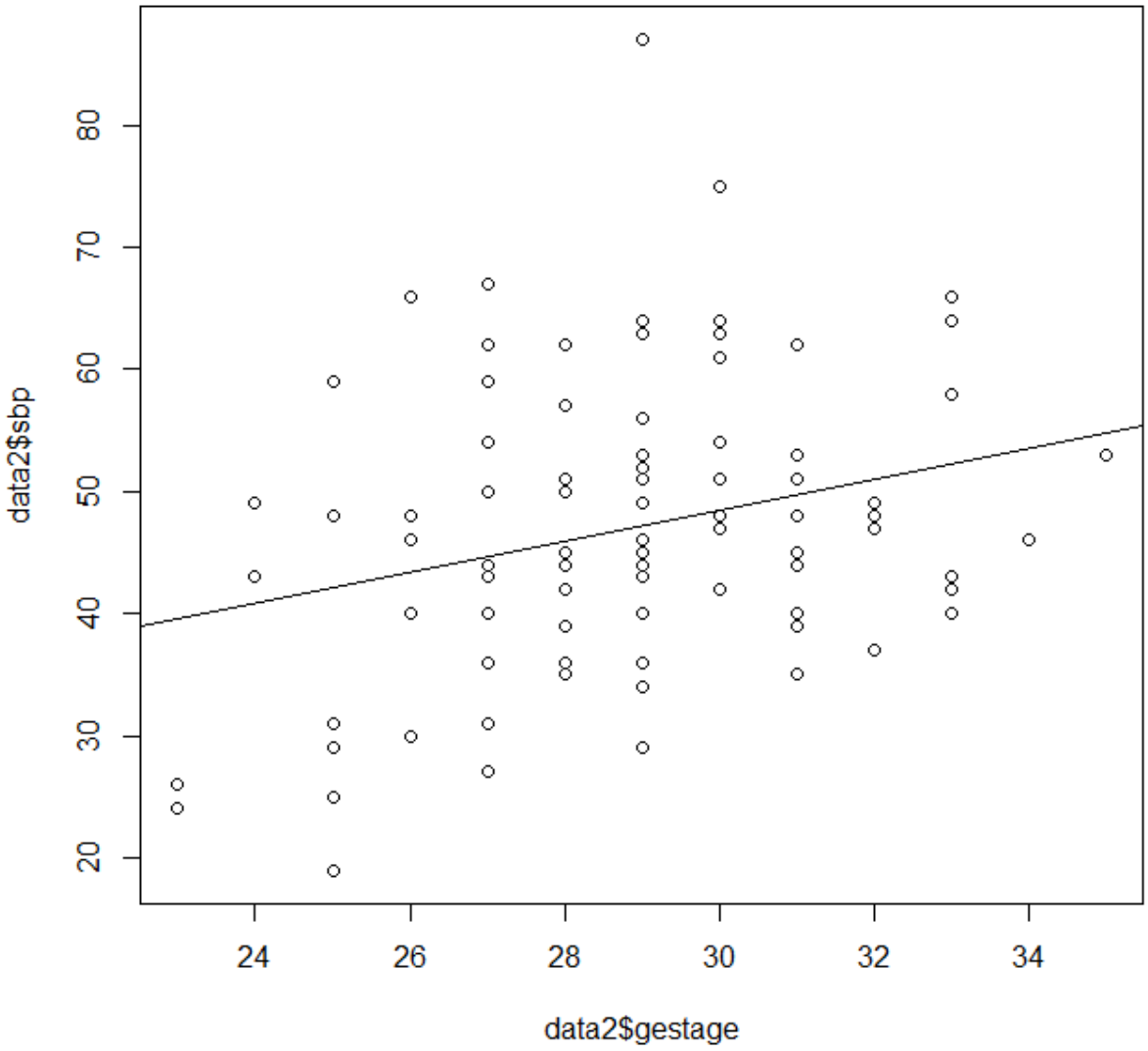
e. (5 points) What is the t test statistic for the slope $\beta_1$ and its p-value? What is the F test statistic and its degrees of freedom in the ANOVA table? What's the relationship between the t test and F test here?

Based on the regression model summary, the t-test statistic for the slope $\beta_1$ is 2.381 and its p-value is 0.02121. The F-test statistic is 5.668 and its degrees of freedom is 49. These results test different hypotheses but are used mathematically together. The square of the t-test statistic (2.381^2 = 5.668) is equal to the F-test statistic.

3. (23 points) The data set lowbwt.sas7bdat contains information for a sample of 100 low birth weight infants born in two teaching hospitals in Boston. Measurements of systolic blood pressure are saved under the variable namesbp, and values of gestational age under the variable name gestage.

a. (3 points) Use appropriate graphic method to explore the relationship between systolic blood pressure and gestational age. Does the graph suggest anything about the relationship between these variables?

```
#import data
data2 <- read.csv('C:\\Users\\acrot\\Downloads\\lowbwt.csv')

#view structure of data
str(data2)
head(data2)

#Create a scatter plot for gestage and sbp variables
plot(x = data2$gestage, y = data2$sbp)
#Add line of best fit to scatter plot
abline(lm(data2$sbp ~ data2$gestage))
```




Based on the line of best fit, there appears to be an increase in systolic blood pressure as the gestational age increases. This appears to align with research on neonates and blood pressure in existing literature.

b. (4 points) Using systolic blood pressure as the response variable and gestational age as the predictor variable, compute the least squares regression line. Interpret the slope and the intercept of the line?

```
#use method of least squares to fit regression line
model2 <- lm(data2$sbp ~ data2$gestage)

#view regression model summary
summary(model2)
```

```
Call:
lm(formula = data2$sbp ~ data2$gestage)

Residuals:
    Min      1Q  Median      3Q     Max
-23.162  -7.828  -1.483   5.568  39.781

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    10.5521    12.6506   0.834  0.40625
data2$gestage   1.2644     0.4362   2.898  0.00463 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11 on 98 degrees of freedom
Multiple R-squared:  0.07895,   Adjusted R-squared:  0.06956
F-statistic: 8.401 on 1 and 98 DF,  p-value: 0.004628
```

The least-squares regression line is 10.5521 + 1.2644(gestational age). The estimated slope is 10.5521, and the intercept is 1.2644. This means that for someone with a gestational age of 0 weeks, they can theoretically expect a systolic blood pressure of 10.5521. This is illogical, as at that point, the organs have not begun to form. For each additional week of gestation, they can expect the systolic blood pressure to increase by 1.2644.

c. (4 points) At the 0.05 level of significance, test the null hypothesis that the population slope $\beta_1$ is equal to 0. What do you conclude?

Based on the regression model summary, the p-value for the slope is 0.40625. This is greater than the significance level of 0.05. This means we fail to reject the null hypothesis of there being no linear relationship between the variables. So we conclude there is no linear relationship between the variables in this population.

d. (4 points) What is the estimated mean systolic blood pressure for the group of infants whose gestational age is 31 weeks? Construct a 95% confidence interval for the true mean value of systolic blood pressure when X = 31 weeks.

```
#Create a data subset where gestational age is 31 weeks
subset_data <- subset(data2, data2$gestage == 31)

#Calculate the mean systolic blood pressure for the subset of data
mean_systolic_bp <- mean(subset_data$sbp)

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

#Print the estimated mean systolic blood pressure and the confidence interval
print(mean_systolic_bp)
print(confidence_interval)
```

```
> print(mean_systolic_bp)
[1] 46.63636
> print(confidence_interval)
[1] 41.57971 51.69302
attr(,"conf.level")
[1] 0.95
>
```

The estimated mean systolic blood pressure for the group of infants whose gestational age is 31 weeks is 46.63636. The 95% confidence interval for the true mean value of systolic blood pressure at 31 weeks is (41.57971, 51.59302).

e. (5 points) Suppose that you randomly select a new child from the population of low birth weight infants and find that his or her gestational age is 31 weeks. What is the predicted systolic blood pressure for this child? Construct a 95% prediction interval for this new value of systolic blood pressure.

```
# Create new data frame with the desired gestational age value
newdatabp2 <- data.frame(gestage = 31)

# Generate predictions and prediction interval
prediction <- predict(model2, newdata = newdatabp2)

#Construct the 95% prediction interval
prediction_interval <- predict(model2, newdata = newdatabp2, interval = "prediction")

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

```
> print(prediction)
       1        2        3        4        5        6        7        8        9       10       11       12       13
47.21908 49.74784 52.27660 49.74784 48.48346 42.16156 44.69032 47.21908 45.95470 47.21908 43.42594 48.48346 47.21908
      14       15       16       17       18       19       20       21       22       23       24       25       26
47.21908 47.21908 47.21908 47.21908 52.27660 52.27660 47.21908 45.95470 48.48346 44.69032 52.27660 51.01222 45.95470
      27       28       29       30       31       32       33       34       35       36       37       38       39
47.21908 45.95470 47.21908 48.48346 49.74784 48.48346 49.74784 47.21908 44.69032 44.69032 44.69032 51.01222 49.74784
      40       41       42       43       44       45       46       47       48       49       50       51       52
45.95470 48.48346 47.21908 45.95470 49.74784 44.69032 42.16156 48.48346 45.95470 45.95470 42.16156 39.63280 44.69032
      53       54       55       56       57       58       59       60       61       62       63       64       65
45.95470 44.69032 44.69032 43.42594 42.16156 39.63280 43.42594 40.89718 47.21908 47.21908 44.69032 48.48346 48.48346
      66       67       68       69       70       71       72       73       74       75       76       77       78
51.01222 52.27660 44.69032 49.74784 43.42594 44.69032 44.69032 54.80536 45.95470 48.48346 49.74784 48.48346 44.69032
      79       80       81       82       83       84       85       86       87       88       89       90       91
42.16156 42.16156 43.42594 47.21908 47.21908 53.54098 48.48346 47.21908 52.27660 48.48346 47.21908 40.89718 52.27660
      92       93       94       95       96       97       98       99      100
42.16156 51.01222 49.74784 49.74784 49.74784 47.21908 51.01222 52.27660 45.95470
> print(prediction_interval)
        fit      lwr      upr
1  47.21908 25.28183 69.15633
```

The predicted systolic blood pressure for this child is 47.21908. The 95% confidence interval for this new value is (25.28183, 69.15633).

f. (3 points) Does the least squares regression model seem to fit the observed data? Comment on the coefficient of determination.

The R-squared value, which represents the coefficient of determination, of the least squares regression model is 0.07895. A higher R-squared value suggests a better fit of the model to the observed data. This is a rather low value, indicating a poor fit of the model to the observed data. This means that the variance in the response variable is not well explained by the explanatory variable.