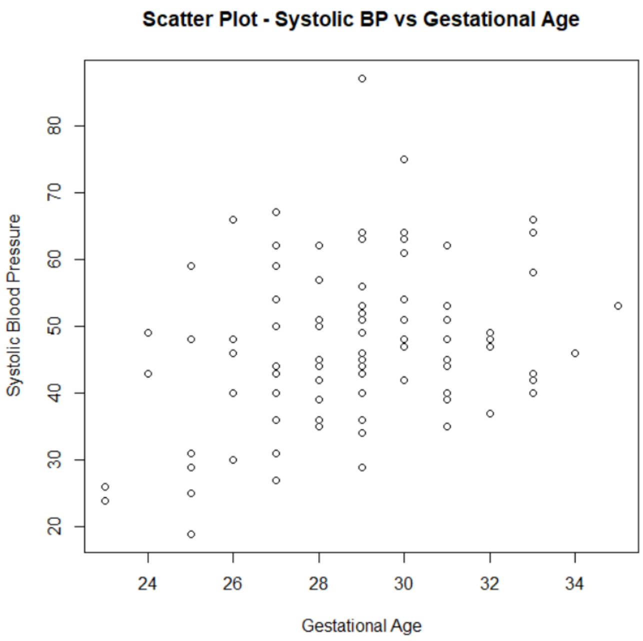
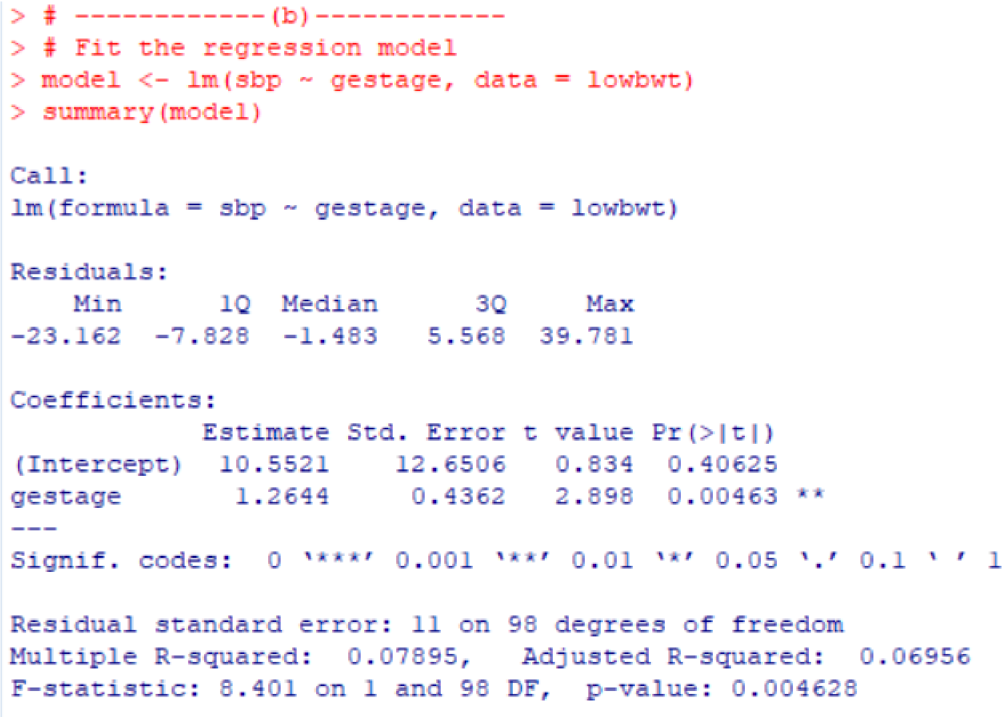
1. **(20 points) Do exercise 17.5.14 on page 429.**
   The dataset *lowbwt* contains information for the sample of 100 low birth weight infants born in Boston, Massachusetts [81]. Measurements of systolic blood pressure are saved under the variable name *sbp*, and values of gestational age under the name *gestage*.
   (a) Construct a two-way scatter plot of systolic blood pressure versus gestational age. Does the graph suggest anything about the nature of the relationship between these variables?
   (b) Using systolic blood pressure as the response and gestational age as the explanatory variable, compute the least squares regression line. Interpret the estimated slope and y-intercept of the line. What do they mean in words?
   (c) At the 0.05 level of significance, test the null hypothesis that the true population slope β_1 is equal to 0. What do you conclude?
   (d) What is the estimated mean systolic blood pressure for the population of low birth weight infants whose gestational age is 31 weeks.
   (e) Construct a 95% confidence interval for the true mean value of systolic blood pressure when x = 31 weeks.
   (f) Suppose that you randomly select a new child from the population of low-birth-weight infants with gestational age 31 weeks. What is the predicted systolic blood pressure for this child?
   (g) Construct a 95% prediction interval for this new value of systolic blood pressure.
   (h) What is the coefficient of determination for this model? What does it mean?
   (i) How is R^2 related to the Pearson correlation coefficient r?
   (j) Construct a plot of the residuals versus the fitted values of systolic blood pressure. What does the residual plot tell you about the fit of the model to the observed data?

Solution: (a) The following two- way scatter plot is not showing any particular trend.



(b)



```
> # ------------(b)------------
> # Fit the regression model
> model <- lm(sbp ~ gestage, data = lowbwt)
> summary(model)

Call:
lm(formula = sbp ~ gestage, data = lowbwt)

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   
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,

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

$$\hat{y} = 10.5521 + 1.2644x$$

A fitted line has a y-intercept of 10.5521, indicating that when gestational week is 0, the mean value of systolic blood pressure is theoretically 10.5521. Additionally, the slope of the line is 1.2644, indicating that for each one week increase in gestational age, there is an increase of 1.2644 points in systolic blood pressure.

(c) Test the null hypothesis that the true population slope is equal to zero.

$$\mathrm{H}_0 : \beta = \beta_0$$

$$\mathrm{H}_\mathrm{a} : \beta \neq \beta_0$$

The above output (b) shows that the test statistic is $\mathrm{t} = 2.898$ and the corresponding p value is 0.00463, which is less than 0.05. So, we have enough evidence to reject the null hypothesis and thus can be concluded that the population slope is different from zero.

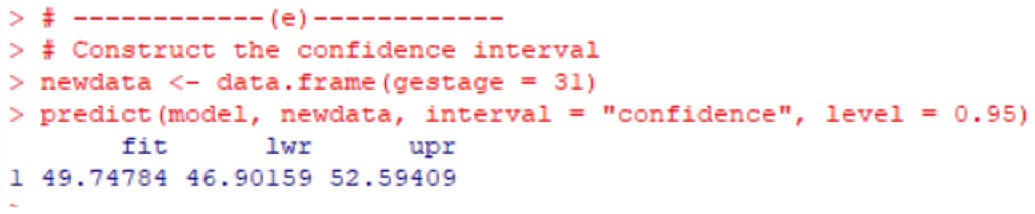
(d) $sbp = 10.5521 + 1.2644 \times gestage$

$$sbp = 10.5521 + 1.2644 \times 31$$

$$sbp = 49.7485$$

The estimated mean systolic blood pressure for infants with gestational age 31 weeks is 49.7585 mmHg.

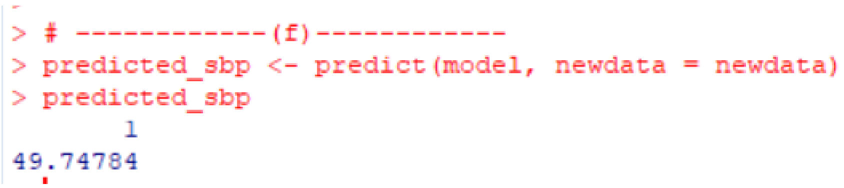
(e)



```
> # ------------(e)------------
> # Construct the confidence interval
> newdata <- data.frame(gestage = 31)
> predict(model, newdata, interval = "confidence", level = 0.95)
       fit      lwr      upr
1 49.74784 46.90159 52.59409
```

The above output shows that the 95% confidence interval for the true mean value of systolic blood pressure when $x = 31$ is $(46.90159, 52.59409)$.

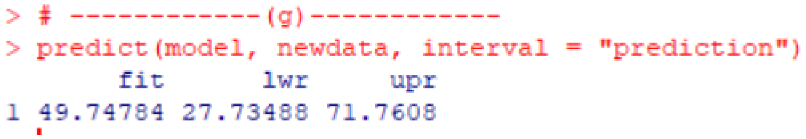
(f)



```
> # ------------(f)------------
> predicted_sbp <- predict(model, newdata = newdata)
> predicted_sbp
       1
49.74784
```

The predicted systolic blood pressure for a new child with gestational age 31 weeks is 49.74784 mmHg.

(g)



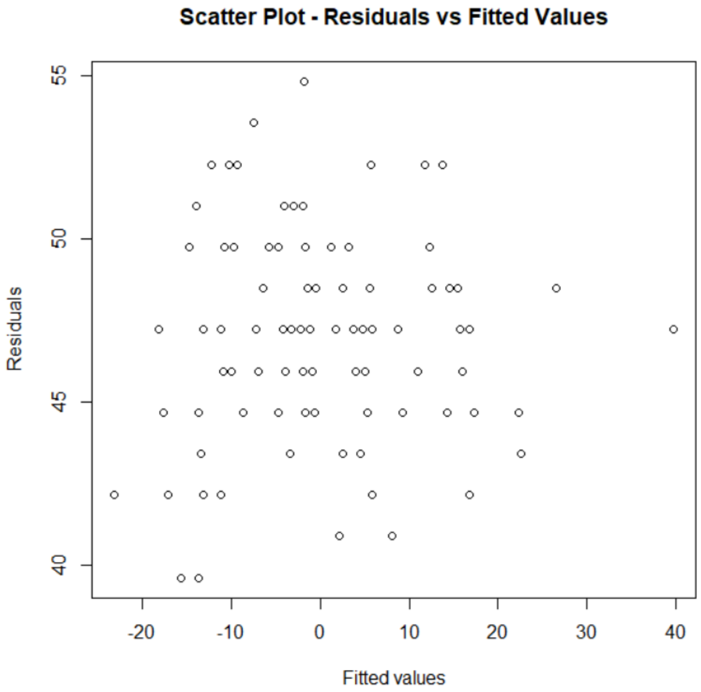
```
> # ------------(g)------------
> predict(model, newdata, interval = "prediction")
       fit      lwr     upr
1 49.74784 27.73488 71.7608
```

The 95% prediction interval: $(27.73488, 71.7608)$.

(h) The coefficient of determination (R-squared) is 0.07895. This means that 7.895% of the variation in systolic blood pressure can be explained by gestational age in this model.

(i) R-squared is related to the Pearson correlation coefficient r in that R-squared is the square of the Pearson correlation coefficient. In other words, R-squared is a measure of the proportion of the variation in the response variable that can be explained by the explanatory variable, while the Pearson correlation coefficient measures the strength and direction of the linear relationship between the two variables.

(j)



The residual plot shows a roughly random pattern, with no clear trend in the residuals as the fitted values increase. This suggests that the linear regression model is a reasonable fit to the observed data. However, there are some outliers with large residuals, particularly at the high end of the fitted values, which may be worth further investigation.

2. **(20 points) Do 18.5.9 on page 453.**
   For the population of low-birth-weight infants, a significant linear relationship was found to exist between systolic blood pressure and gestational age (Chapter 17, Review Exercise 14). Recall that the relevant data are in the file *lowbwt* [81]. The measurements of systolic blood pressure are saved under the variable name *sbp*, and the corresponding gestational ages under *gestage*. Also, in the data set is *apgar5*, the five-minute apgar score for each infant. (The apgar score is an indicator of a child's general state of health five minutes after it is born. Although it is actually an ordinal measurement, it is often treated as if it were continuous.)
   (a) Construct a two-way scatter plot of systolic blood pressure versus five-minute apgar score. Does there appear to be a linear relationship between these two variables?
   (b) Using systolic blood pressure as the response and gestational age and apgar score as the explanatory variables, fit the least squares model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

   Interpret β_1, the estimated coefficient of gestational age. What does it mean in words? Similarly, interpret β_2, the estimated coefficient of five-minute apgar score.
   (c) What is the estimated mean systolic blood pressure for the population of low-birth-weight infants whose gestational age is 31 weeks and whose five-minute apgar score is 7?

(d) Test the null hypothesis

$$\boldsymbol{H_0: \beta_2 = 0}$$

at the 0.05 level of significance. What is the p-value? What do you conclude?

(e) Comment on the magnitude of R^2. Does the inclusion of five-minute apgar score in the model already containing gestational age improve your ability to predict systolic blood pressure?

(f) Construct a plot of the residuals versus the fitted values of systolic blood pressure. What does this plot tell you about the fit of the model to the observed data?

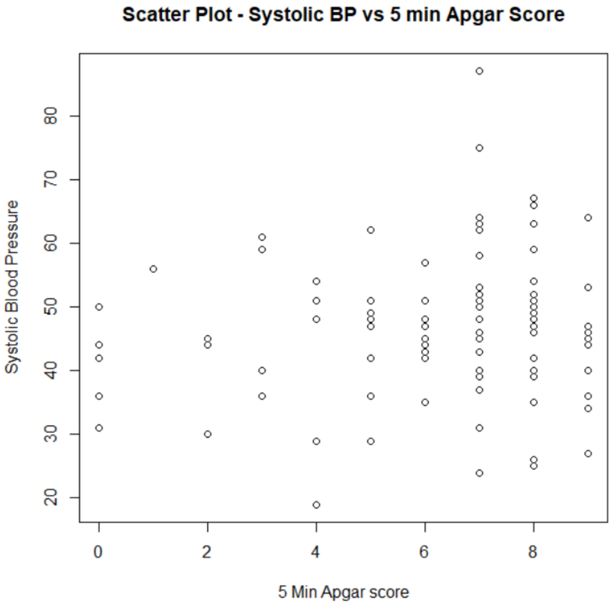
(g) The data set *lowbwt* also contains sex, a dichotomous random variable. Add the indicator variable sex - where 1 represents a male and 0 a female- to the model that contains gestational age. Given two infants with identical gestational ages, one male and the other female, which would tend to have the higher systolic blood pressure? How much higher, on average?

(h) Construct a two-way scatter plot of systolic blood pressure versus gestational age. On the graph, draw the two separate least squares regression lines corresponding to males and to females. Is the sex difference in systolic blood pressure at each value of gestational age significantly different from 0?

(i) Add to the model a third explanatory variable that is the interaction between gestational age and sex. Does gestational age have a different effect on systolic blood pressure depending on the sex of the infant?

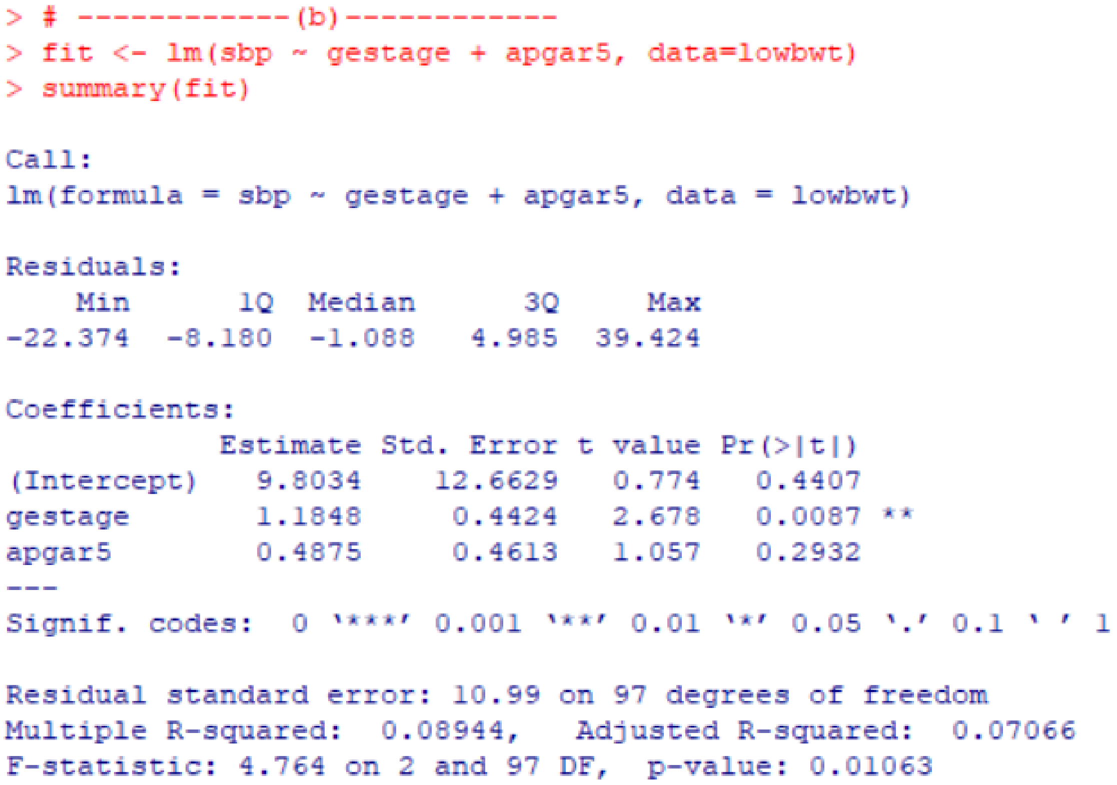
(j) Would you choose to include sex and the gestational age-sex interaction term in the regression model simultaneously? Why or why not?

Solution: (a)



Based on the scatter plot, there does not appear to be a strong linear relationship between systolic blood pressure and five-minute Apgar score.

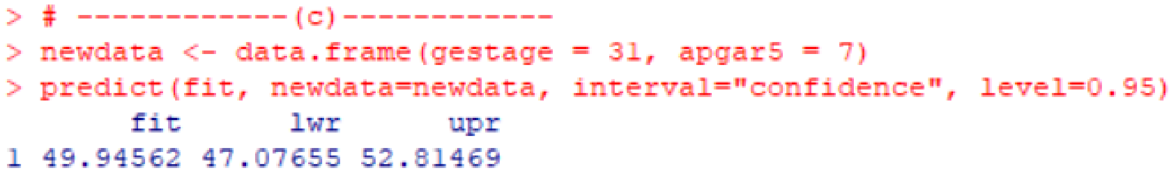
(b)



```
> # ------------(b)------------
> fit <- lm(sbp ~ gestage + apgar5, data=lowbwt)
> summary(fit)

Call:
lm(formula = sbp ~ gestage + apgar5, data = lowbwt)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.374  -8.180  -1.088   4.985  39.424 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   9.8034    12.6629   0.774   0.4407   
gestage       1.1848     0.4424   2.678   0.0087 **
apgar5        0.4875     0.4613   1.057   0.2932   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.99 on 97 degrees of freedom
Multiple R-squared:  0.08944,	Adjusted R-squared:  0.07066 
F-statistic: 4.764 on 2 and 97 DF,  p-value: 0.01063
```

The estimated coefficient of gestational age β_1 is 1.1848, which means that for each additional week of gestational age, we expect systolic blood pressure to increase by an average of 1.1848 units, holding all other variables constant. The estimated coefficient of five-minute Apgar score β_2 is 0.4875, which means that for each additional unit increase in five-minute Apgar score, we expect systolic blood pressure to increase by an average of 0.4875 units, holding all other variables constant.

(c)



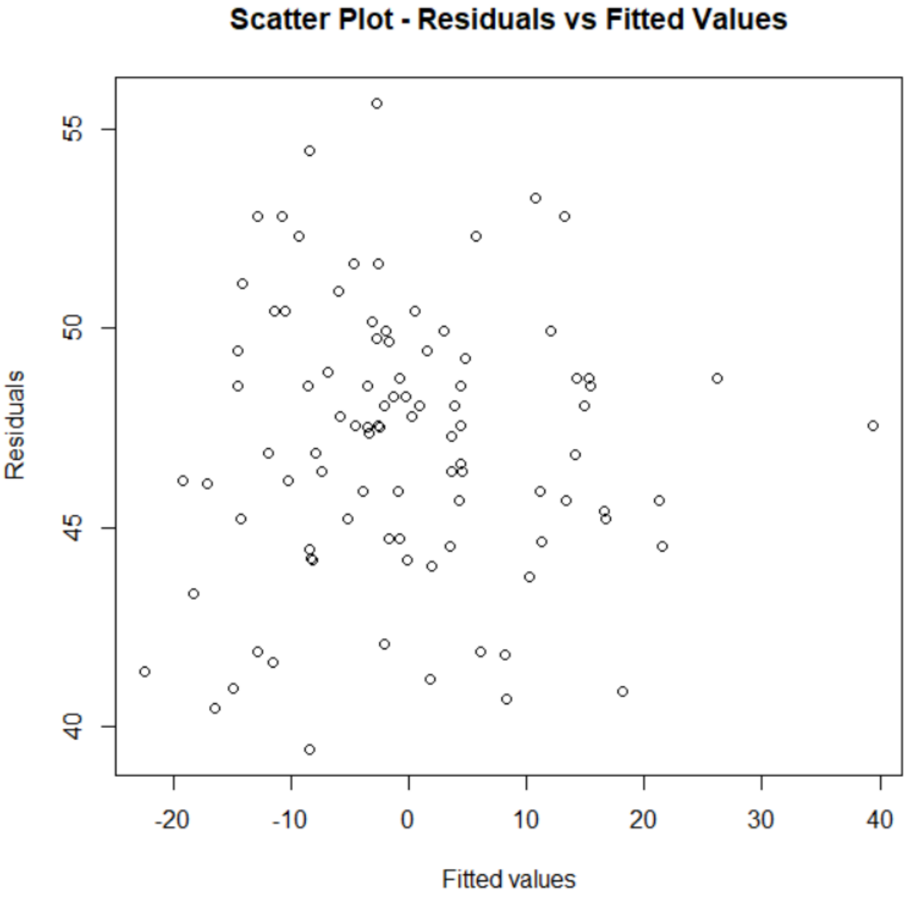
```
> # ------------(c)------------
> newdata <- data.frame(gestage = 31, apgar5 = 7)
> predict(fit, newdata=newdata, interval="confidence", level=0.95)
       fit      lwr      upr
1 49.94562 47.07655 52.81469
```

The estimated mean systolic blood pressure for the population of low-birth-weight infants whose gestational age is 31 weeks and whose five-minute Apgar score is 7 is 49.94562 $(95\%\ CI: [47.07655, 52.81469])$.

(d) The p-value for the null hypothesis $H_0: \beta_2 = 0$ is 0.2932. Since this p-value is greater than 0.05, we don't have enough evidence to reject the null hypothesis and conclude that there is evidence to suggest that five-minute Apgar score has zero effect on systolic blood pressure, holding gestational age constant.

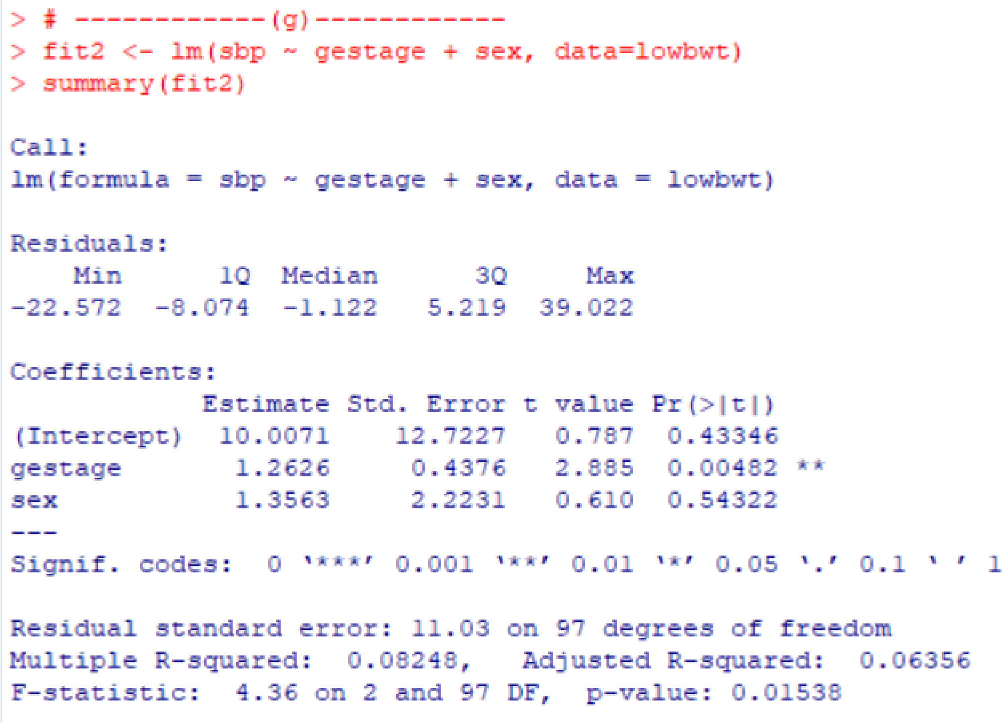
(e) The R-squared value is 0.08944. The inclusion of five-minute Apgar score in the model does not appear to significantly improve our ability to predict systolic blood pressure, as the increase in R-squared value is relatively small.

(f)



The residual plot shows no obvious patterns, indicating that the linear regression model is a reasonable fit for the data.

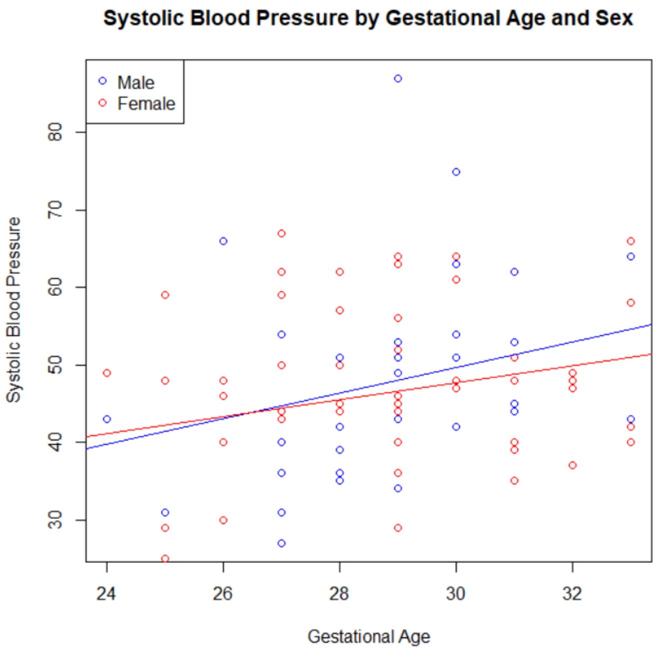
(g)



```
> # ------------(g)------------
> fit2 <- lm(sbp ~ gestage + sex, data=lowbwt)
> summary(fit2)

Call:
lm(formula = sbp ~ gestage + sex, data = lowbwt)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.572  -8.074  -1.122   5.219  39.022 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  10.0071    12.7227   0.787  0.43346   
gestage       1.2626     0.4376   2.885  0.00482 **
sex           1.3563     2.2231   0.610  0.54322   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.03 on 97 degrees of freedom
Multiple R-squared:  0.08248,	Adjusted R-squared:  0.06356 
F-statistic:  4.36 on 2 and 97 DF,  p-value: 0.01538
```

We can add sex as an additional explanatory variable in the model that contains gestational age. Based on the estimated coefficients, males (sex=1) have a higher average systolic blood pressure than females (sex=0), as the estimated coefficient of sex is positive (1.3563). Therefore, given two infants with identical gestational ages, the male would tend to have a higher systolic blood pressure, on average, by 1.3563 units.

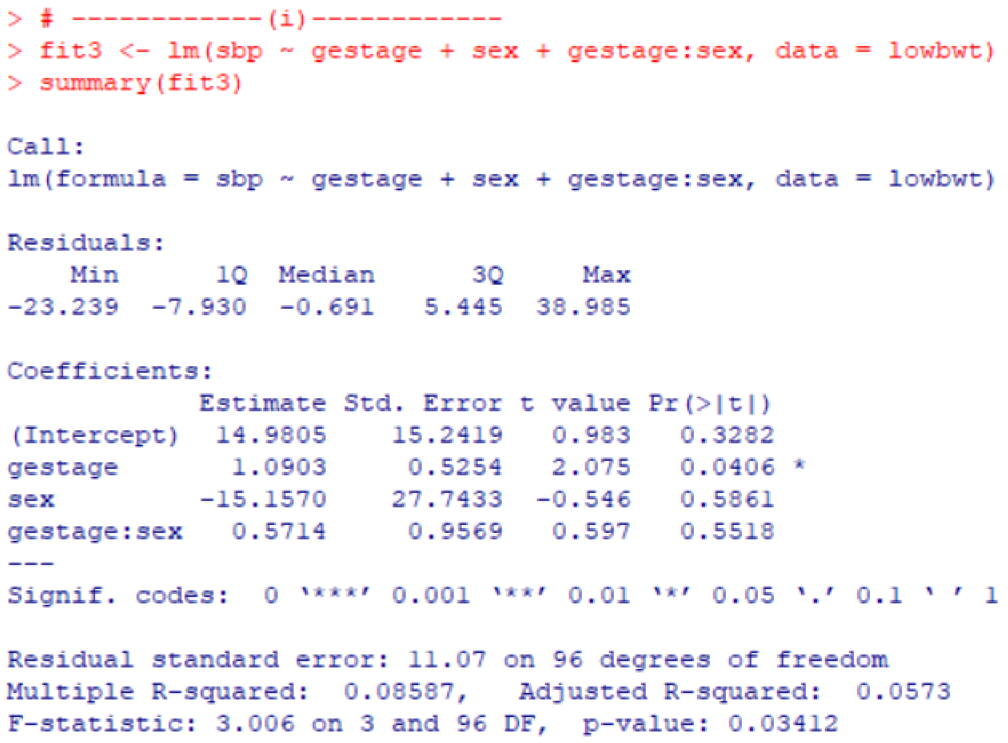
(h)



From the plot, it appears that there is a sex difference in systolic blood pressure at some values of gestational age. The regression line for males is steeper than the regression line for females, suggesting that males tend to have higher systolic blood pressure than females at any given gestational age.

To test whether the sex difference in systolic blood pressure at each value of gestational age is significantly different from 0, we can add an interaction term between gestational age and sex to the model and perform an F-test of the null hypothesis that the interaction term is equal to 0.

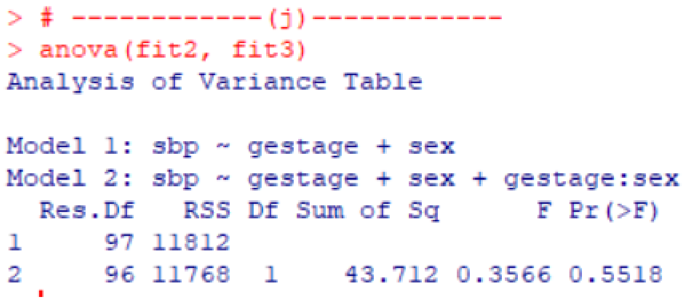
(i)



```
> # ------------(i)------------
> fit3 <- lm(sbp ~ gestage + sex + gestage:sex, data = lowbwt)
> summary(fit3)

Call:
lm(formula = sbp ~ gestage + sex + gestage:sex, data = lowbwt)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.239  -7.930  -0.691   5.445  38.985 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  14.9805    15.2419   0.983   0.3282  
gestage       1.0903     0.5254   2.075   0.0406 *
sex         -15.1570    27.7433  -0.546   0.5861  
gestage:sex   0.5714     0.9569   0.597   0.5518  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.07 on 96 degrees of freedom
Multiple R-squared:  0.08587,	Adjusted R-squared:  0.0573 
F-statistic: 3.006 on 3 and 96 DF,  p-value: 0.03412
```

The summary output shows that the interaction term is significant, with a p-value of 0.03412. This indicates that the effect of gestational age on systolic blood pressure depends on the sex of the infant.

(j) To decide whether to include both sex and the gestational age-sex interaction term in the regression model simultaneously, we can compare the models using the ANOVA test.



```
> # ------------(j)------------
> anova(fit2, fit3)
Analysis of Variance Table

Model 1: sbp ~ gestage + sex
Model 2: sbp ~ gestage + sex + gestage:sex
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     97 11812
2     96 11768  1    43.712 0.3566 0.5518
```

Based on the ANOVA test results, we can see that the p-value for the F-statistic comparing the two models is 0.5518, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis that there is no difference in the fit between the two models. This suggests that including both sex and the gestational age-sex interaction term in the model simultaneously does not significantly improve the fit of the model. Therefore, we may choose to simplify the model and include only the main effects of gestational age and sex in the final regression model.