# Statistical data analysis, Assignment 7

Name: 徐竣霆
ID: 110060012

## Problem 1

### a

- (i)
  $\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$, where $Cov(X,Y)$ is the covariance between $X$ and $Y$, $\sigma_X$ is the standard deviation of $X$, $sigma_Y$ is the standard deviation of $Y$.

- (ii)
  $\gamma = \frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}}$, where $\bar{x}$ is the mean of $X$, $\bar{y}$ is the mean of $Y$.

- (iii)
  $R^2 = \gamma^2$

### b

- (i)
  $Y = \beta_0 + \beta_1 X + \epsilon$, where $\beta_0$ is the intercept parameter, $\beta_1$ is the slope parameter, $\epsilon$ is the error($N(0, \sigma^2)$).

- (ii)
  Findint the value of $\beta_0$ and $\beta_1$ that minimize the sum of squared residuals: $\sum\limits_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$. The least squares estimator $\hat{\beta} = (\hat{\beta_0}, \hat{\beta_1})$ are given by:
  $\hat{\beta_1} = \frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}, \hat{\beta_0} = \bar{y} - \beta_1 \bar{x}$

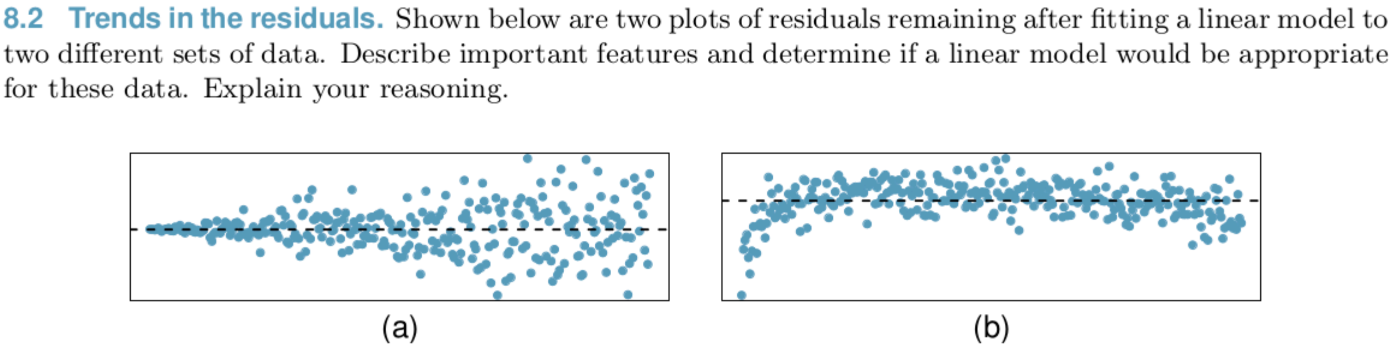
- (iii)
  ANOVA Table:

| Source | SS | DF | MS | F |
|---|---|---|---|---|
| Model | $SSR = SST - SSE$ | $1$ | $MSR = \frac{SST-SSE}{1}$ | $F = \frac{MSR}{MSE}$ |
| Residual | $SSE = \sum (y_i - \hat{y}_i)^2$ | $n-2$ | $MSE = \frac{SSE}{n-2}$ | |
| Total | $SST = \sum (y_i - \bar{y})^2$ | $n-1$ | | |

- (iv)
  $H_0 : \beta_1 = 0$, this hypothesis states that there is no linear relationship between $X$ and $Y$.

# Problem 2

## 8.2

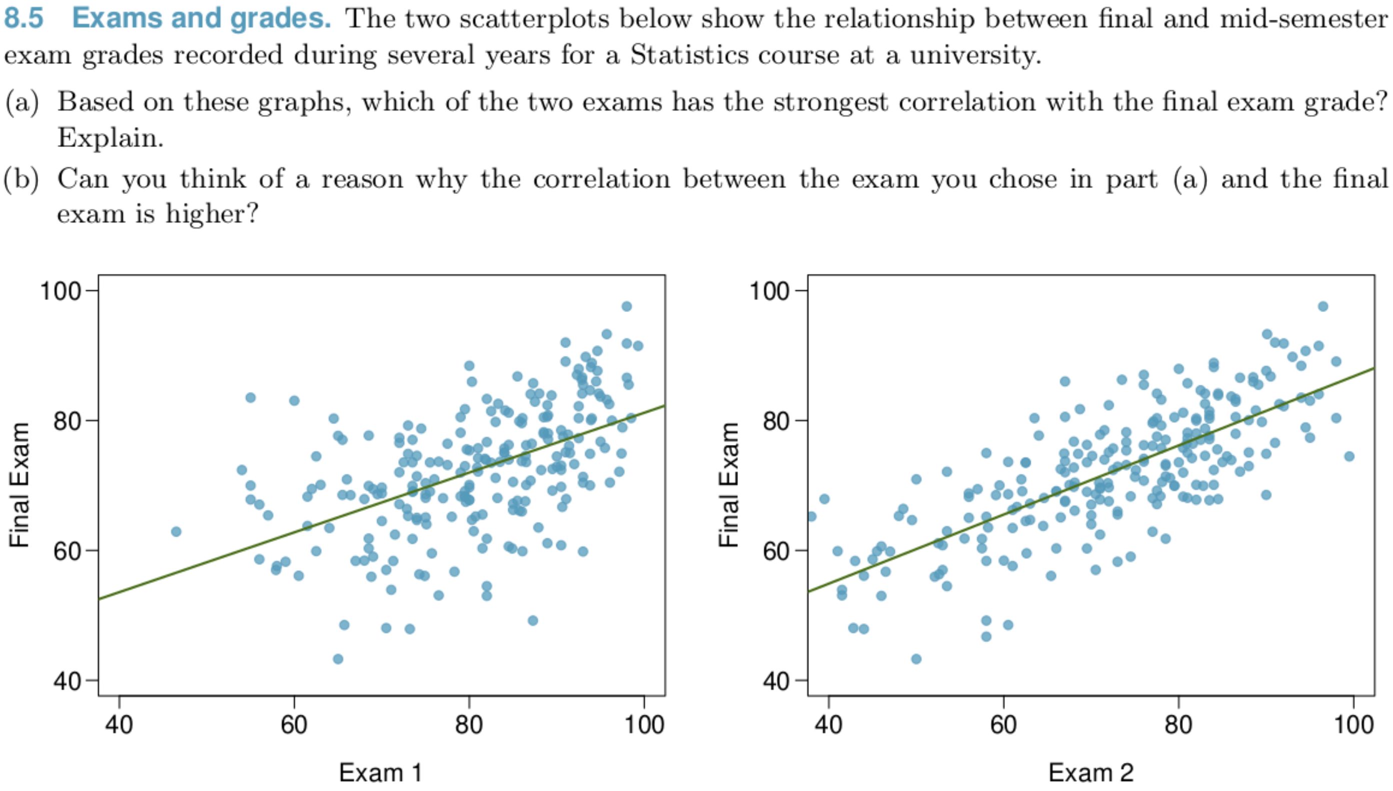


**8.2 Trends in the residuals.** Shown below are two plots of residuals remaining after fitting a linear model to two different sets of data. Describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.

For (a), the variability are higher for larger x, violating the constant variability condition, thus linear model wouldn't be appropriate.

For (b), most of the points are distributed along the line, except for a few at very left which are overestimated(might be outliers), thus a linear model would be appropriate.

## 8.5



**8.5 Exams and grades.** The two scatterplots below show the relationship between final and mid-semester exam grades recorded during several years for a Statistics course at a university.

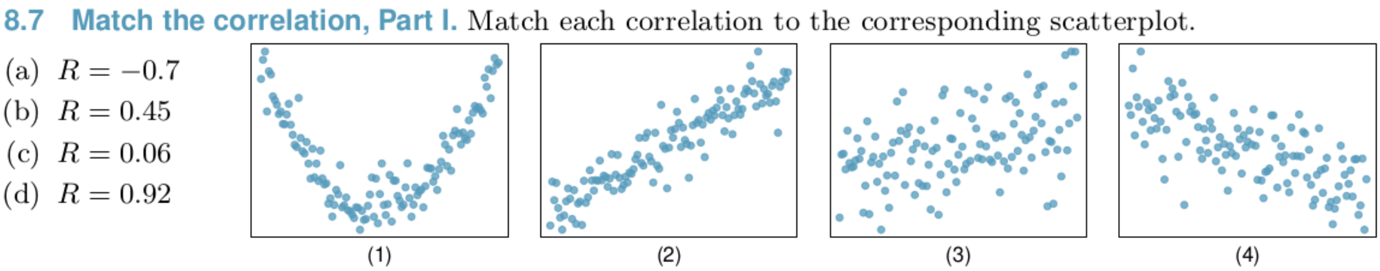
(a) Based on these graphs, which of the two exams has the strongest correlation with the final exam grade? Explain.

(b) Can you think of a reason why the correlation between the exam you chose in part (a) and the final exam is higher?

- (a): Exam 2, since it seems less scattered(the sum of squared residuals seems smaller), so it should have a higher correlation with Final Exam than Exam 1.

- (b): The material that Exam 2 and Final Exam covered might be more similar or even overlapped.

## 8.7

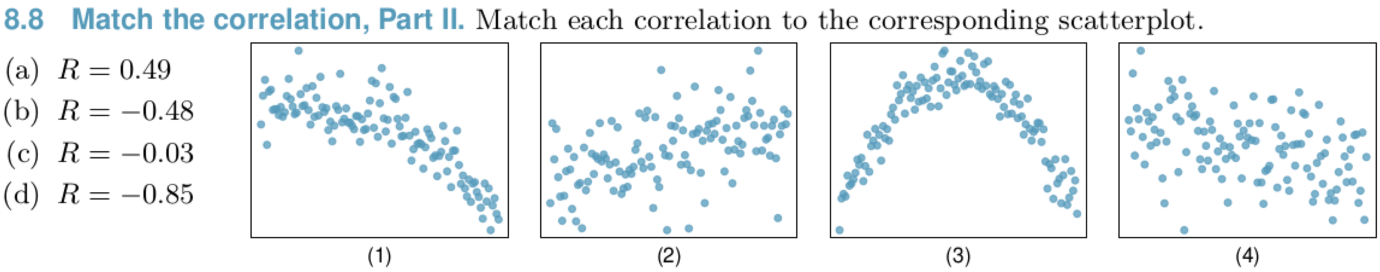


8.7 Match the correlation, Part I. Match each correlation to the corresponding scatterplot.

(a) $R = -0.7$
(b) $R = 0.45$
(c) $R = 0.06$
(d) $R = 0.92$

- (1): c, $R = 0.06$
- (2): d, $R = 0.92$
- (3): b, $R = 0.45$
- (4): a, $R = -0.7$

## 8.8

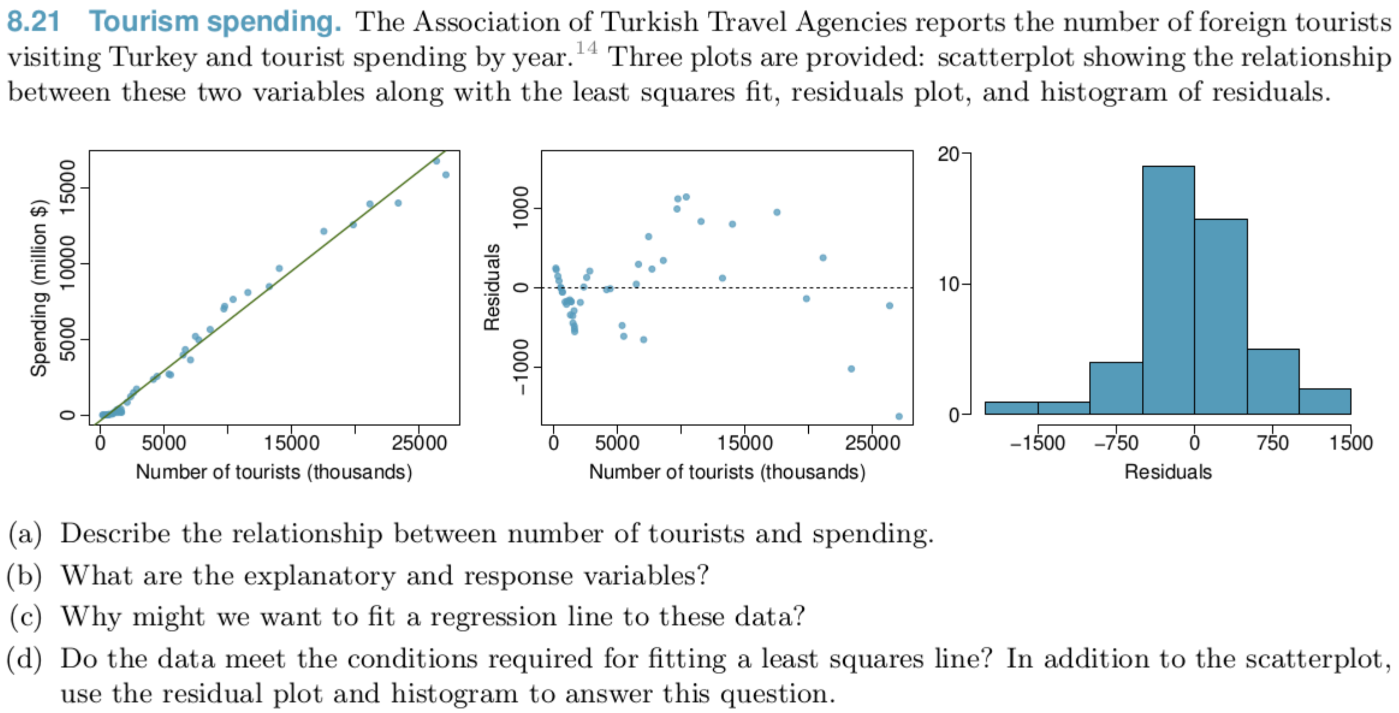


8.8 Match the correlation, Part II. Match each correlation to the corresponding scatterplot.

(a) $R = 0.49$
(b) $R = -0.48$
(c) $R = -0.03$
(d) $R = -0.85$

- (1): d, $R = -0.85$
- (2): a, $R = 0.49$
- (3): c, $R = -0.03$
- (4): b, $R = -0.48$

## 8.21



**8.21 Tourism spending.** The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year.[14] Three plots are provided: scatterplot showing the relationship between these two variables along with the least squares fit, residuals plot, and histogram of residuals.

(a) Describe the relationship between number of tourists and spending.

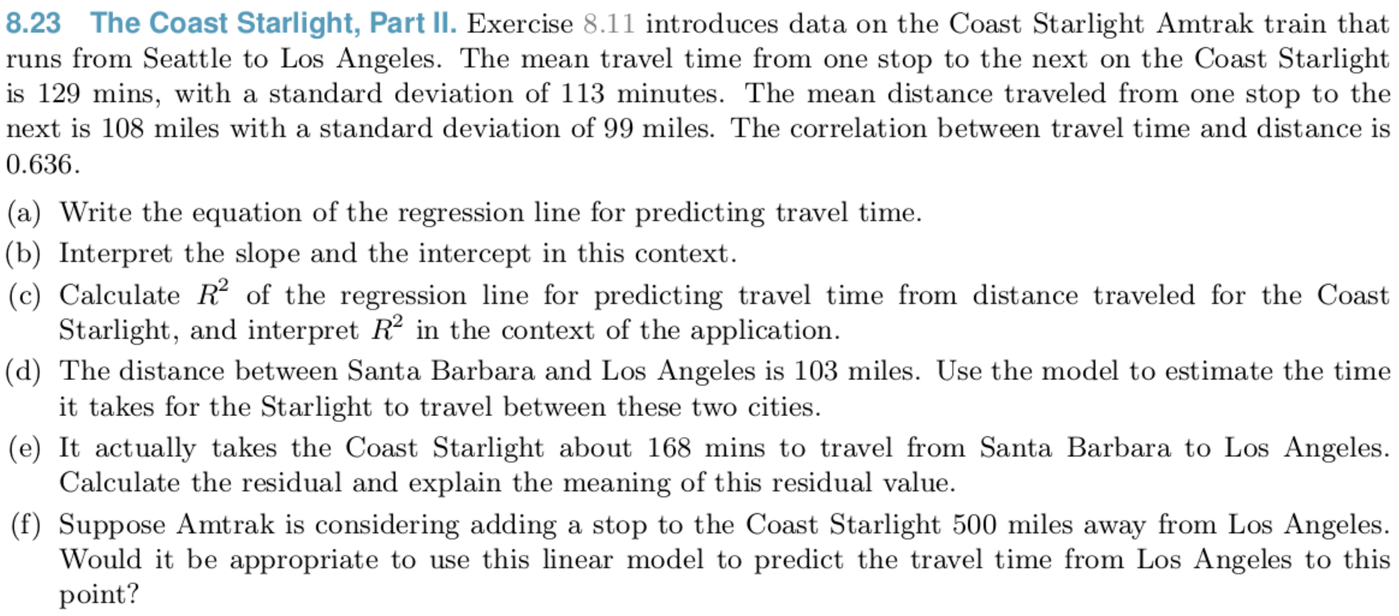
(b) What are the explanatory and response variables?

(c) Why might we want to fit a regression line to these data?

(d) Do the data meet the conditions required for fitting a least squares line? In addition to the scatterplot, use the residual plot and histogram to answer this question.

- a: There is a positive, strong, linear relationship between these two.
- b: The explanatory variable should be Number of Tourists, and the response variable should be Spending.
- c: We can predict spending for a given number of tourists using a regression line.
- d: Although it seems linear on scatterplot(Linearity), and seems normal on histogram(Nearly normal residuals), but it seems violated the Constant variability condition.

## 8.23



**8.23 The Coast Starlight, Part II.** Exercise 8.11 introduces data on the Coast Starlight Amtrak train that runs from Seattle to Los Angeles. The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

(a) Write the equation of the regression line for predicting travel time.

(b) Interpret the slope and the intercept in this context.

(c) Calculate $R^2$ of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret $R^2$ in the context of the application.

(d) The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.

(e) It actually takes the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.

(f) Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

- a:
  $b_1 = \frac{s_y}{s_x} \times R = \frac{113}{99} \times 0.636 = 0.726$
  $b_0 = \bar{y} - b_1 \times \bar{x} = 129 - 0.726 \times 108 = 51$
  $\hat{y} = 51 + 0.726 \times x$, y = travel time, x = distance.

- b:
  $b_1$: For each additional mile in distance, the model predicts an additional $0.726$ minutes in travel time.
  $b_0$: When the distance traveled is $0$ miles, the travel time is expected to be $51$ minutes. It does not make sense to have a travel distance of $0$ miles in this context. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself.

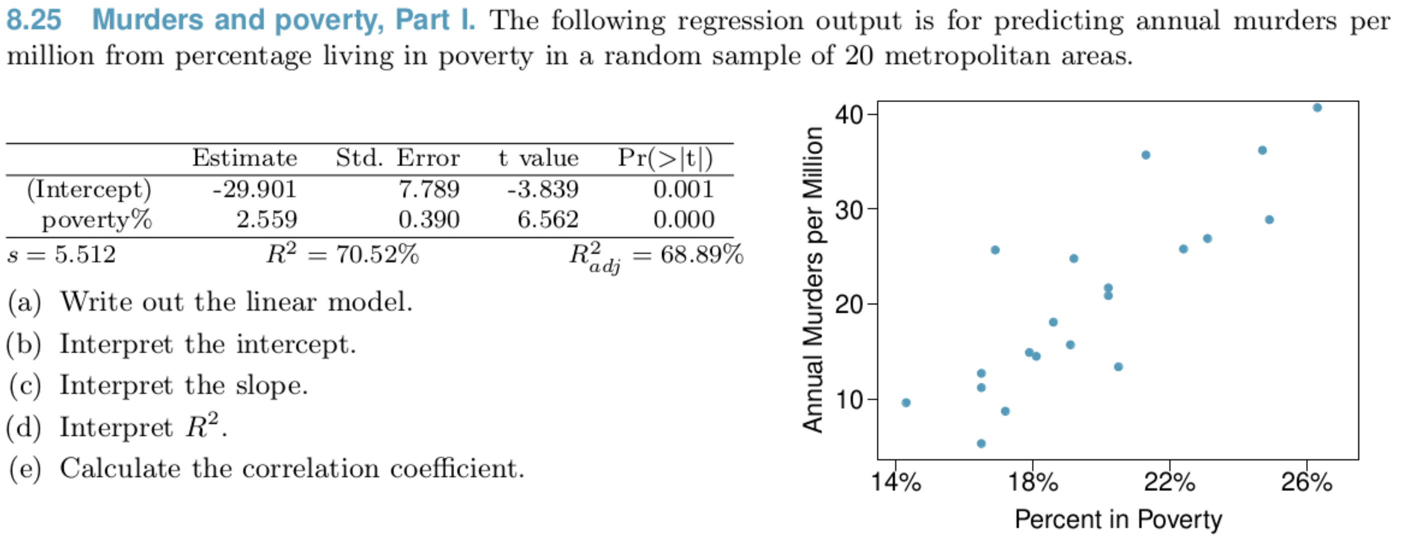
- c:
  $R^2 = 0.636^2 = 0.404$, about $40.4\%$ of the variablity in travel time is explained by the model.

- d:
  $\hat{y} = 51 + 0.726 \times x = 51 + 0.726 \times 103 = 125.778$, roughly $126$ minutes.

- e:
  $e_i = y_i - \hat{y}_i = 168 - 126 = 42$ minutes. A positive residual means the model underestimates the travel time.

- f:
  No, this would require extrapolation, which can be very unreliable.

## 8.25



**8.25 Murders and poverty, Part I.** The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -29.901 | 7.789 | -3.839 | 0.001 |
| poverty% | 2.559 | 0.390 | 6.562 | 0.000 |
| $s = 5.512$ | | $R^2 = 70.52\%$ | | $R^2_{adj} = 68.89\%$ |

(a) Write out the linear model.
(b) Interpret the intercept.
(c) Interpret the slope.
(d) Interpret $R^2$.
(e) Calculate the correlation coefficient.

- a: $\hat{y} = -29.901 + 2.559 \times x$, y is the annual murders per million, x is the percent in poverty.

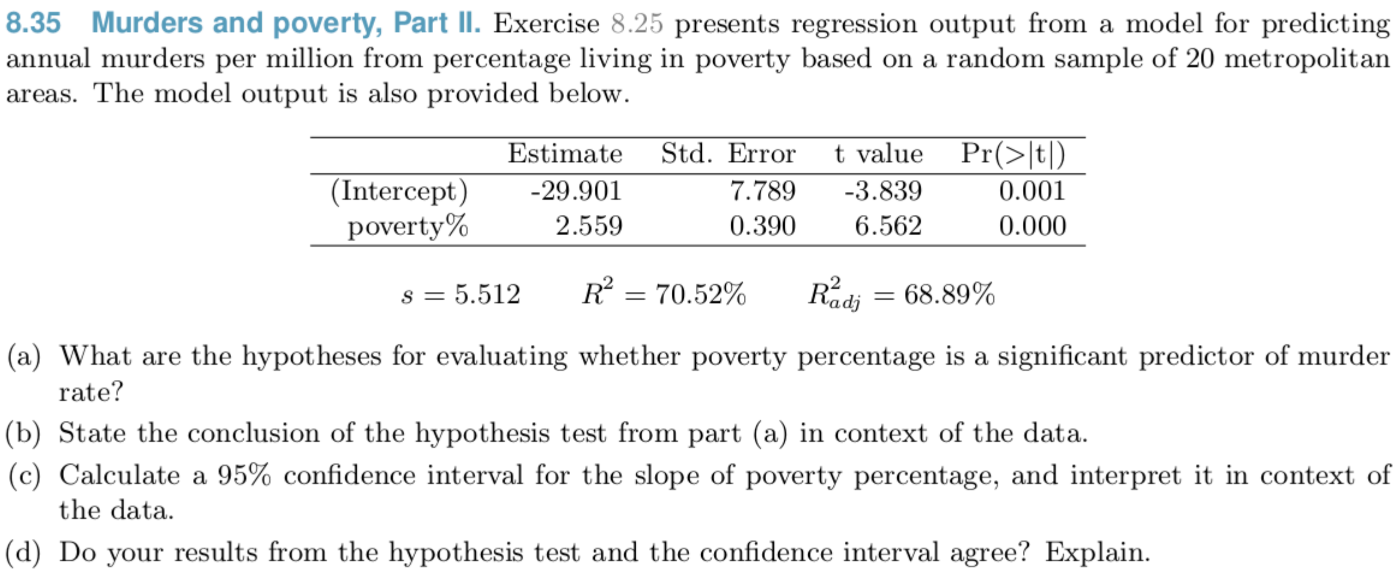
- b: Expected murder rate in metropolitan areas with no poverty is $-29.901$ per million. This has no meanings, it just serves to adjust the height of the regression line.

- c: For each additional percentage increase in poverty, we expect murders per million to be higher on average by $2.559$.

- d: About $70.52\%$ of the variablity of murder rates in metropolitan areas is explained by poverty level.

- e: $\sqrt{0.7052} = 0.8398$

## 8.35



**8.35 Murders and poverty, Part II.** Exercise 8.25 presents regression output from a model for predicting annual murders per million from percentage living in poverty based on a random sample of 20 metropolitan areas. The model output is also provided below.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -29.901 | 7.789 | -3.839 | 0.001 |
| poverty% | 2.559 | 0.390 | 6.562 | 0.000 |

$$s = 5.512 \qquad R^2 = 70.52\% \qquad R^2_{adj} = 68.89\%$$

(a) What are the hypotheses for evaluating whether poverty percentage is a significant predictor of murder rate?

(b) State the conclusion of the hypothesis test from part (a) in context of the data.

(c) Calculate a 95% confidence interval for the slope of poverty percentage, and interpret it in context of the data.

(d) Do your results from the hypothesis test and the confidence interval agree? Explain.

- a: $H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$

- b: The p-value for this test is approximately $0$, smaller than $0.05$, thus we reject $H_0$.

- c: $n = 20, df = 18, T^*_{18} = 2.10$, $2.559 \pm 2.10 \times 0.390 = (1.74, 3.378)$. For each additional percentage increase in poverty, murder rate is expected to be higher on average by $1.74$ to $3.378$ per million.

- d: Yes, we rejected $H_0$ and the confidence interval does not include $0$.

# 8.44

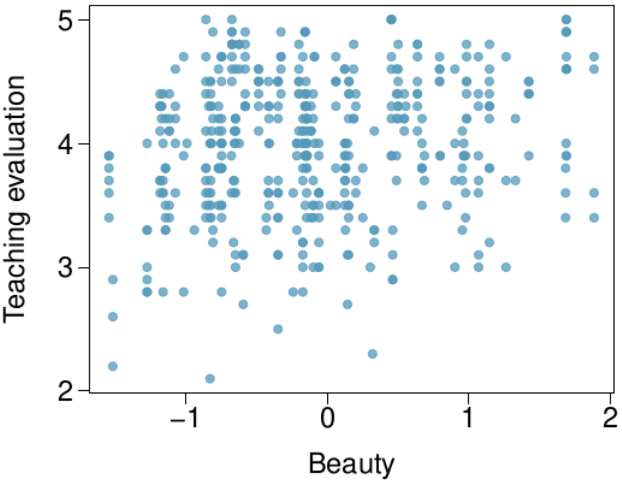
**8.44 Rate my professor.** Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors.[21] The scatterplot below shows the relationship between these variables, and regression output is provided for predicting teaching evaluation score from beauty score.

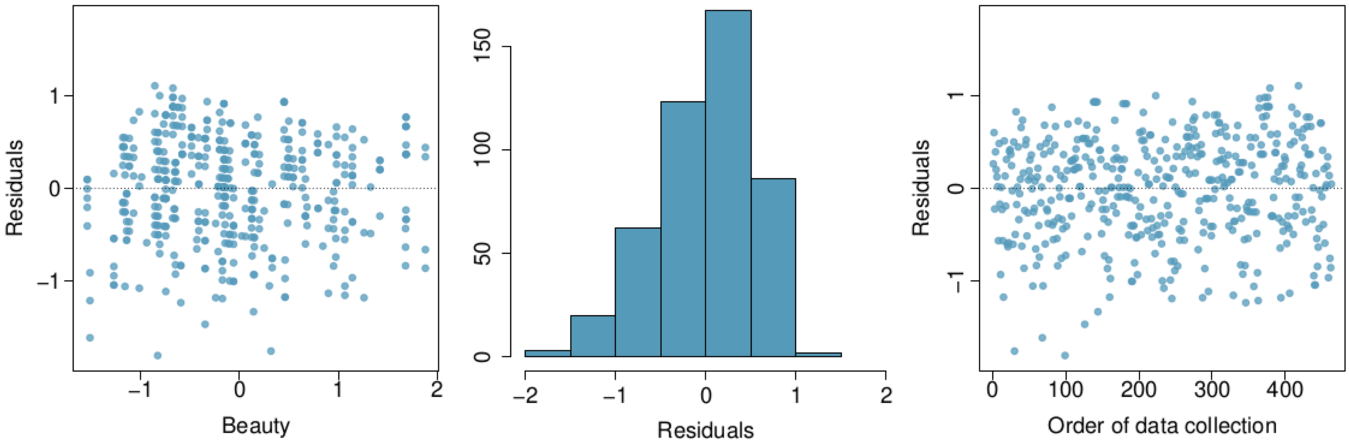
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.010 | 0.0255 | 157.21 | 0.0000 |
| beauty | | 0.0322 | 4.13 | 0.0000 |

(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

(b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

(c) List the conditions required for linear regression and check if each one is satisfied for this model based on the following diagnostic plots.





- a:

$$b_0 = \bar{y} - b_1 \times \bar{x}$$

$$b_1 = \frac{\bar{y}-b_0}{\bar{x}} = \frac{3.9983-4.010}{-0.0883} = 0.1325$$

- b:

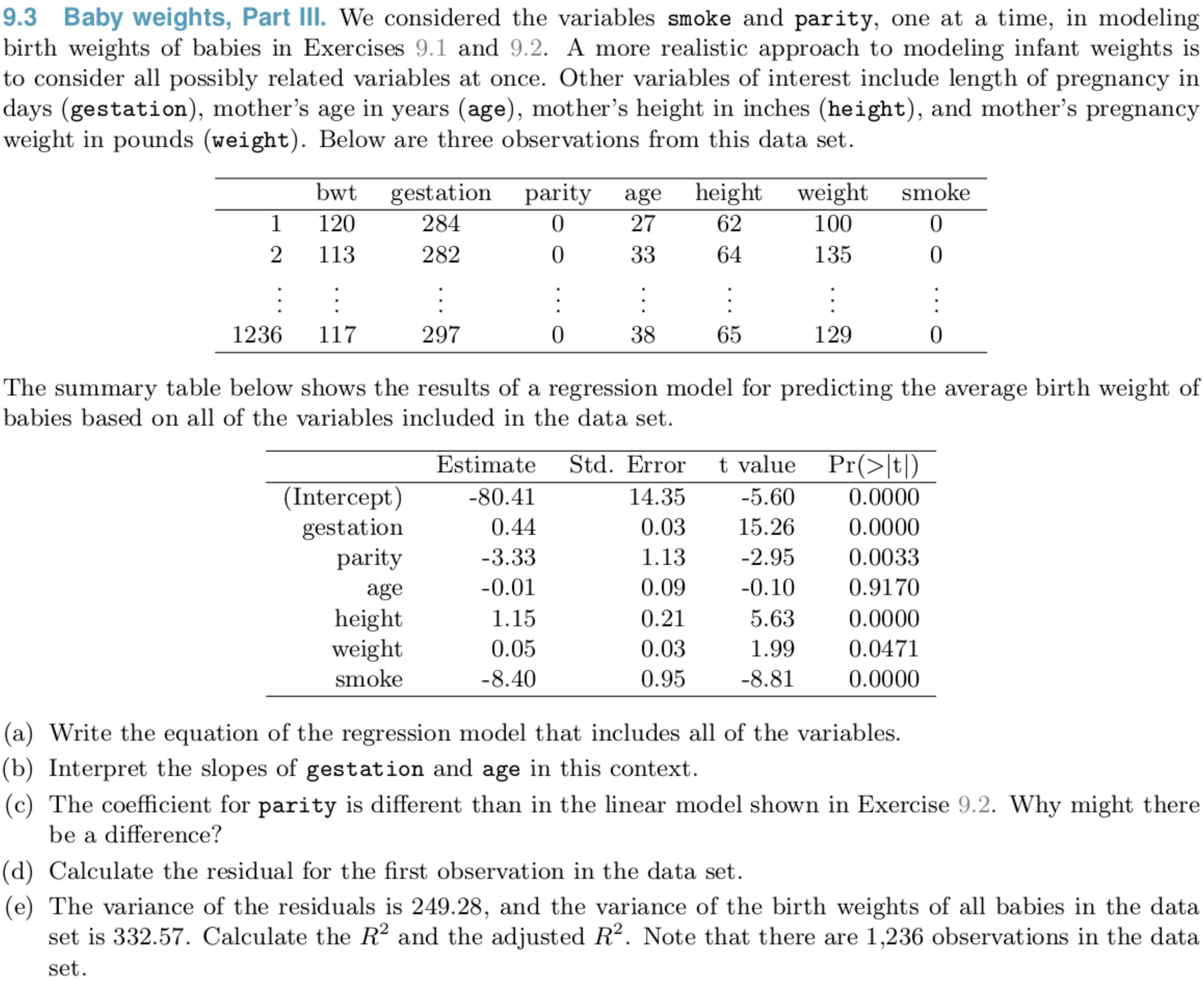
Yes, since the p-value is approximately $0.0000$, smaller than $0.05$, thus we reject $H_0$ that they are not related. And $b_1 = 0.1325 > 0$.

- c:

Linearity: Although $R$ and $R^2$ are not provided, I think the linearity is satisfied(p-value is small).

Nearly normal residuals: As shown in the histogram, they are nearly normal(I think?).
Constant variability: It seems to have constant variability.

## 9.3



**9.3 Baby weights, Part III.** We considered the variables `smoke` and `parity`, one at a time, in modeling birth weights of babies in Exercises 9.1 and 9.2. A more realistic approach to modeling infant weights is to consider all possibly related variables at once. Other variables of interest include length of pregnancy in days (`gestation`), mother's age in years (`age`), mother's height in inches (`height`), and mother's pregnancy weight in pounds (`weight`). Below are three observations from this data set.

| | bwt | gestation | parity | age | height | weight | smoke |
|---|---|---|---|---|---|---|---|
| 1 | 120 | 284 | 0 | 27 | 62 | 100 | 0 |
| 2 | 113 | 282 | 0 | 33 | 64 | 135 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1236 | 117 | 297 | 0 | 38 | 65 | 129 | 0 |

The summary table below shows the results of a regression model for predicting the average birth weight of babies based on all of the variables included in the data set.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -80.41 | 14.35 | -5.60 | 0.0000 |
| gestation | 0.44 | 0.03 | 15.26 | 0.0000 |
| parity | -3.33 | 1.13 | -2.95 | 0.0033 |
| age | -0.01 | 0.09 | -0.10 | 0.9170 |
| height | 1.15 | 0.21 | 5.63 | 0.0000 |
| weight | 0.05 | 0.03 | 1.99 | 0.0471 |
| smoke | -8.40 | 0.95 | -8.81 | 0.0000 |

(a) Write the equation of the regression model that includes all of the variables.

(b) Interpret the slopes of `gestation` and `age` in this context.

(c) The coefficient for `parity` is different than in the linear model shown in Exercise 9.2. Why might there be a difference?

(d) Calculate the residual for the first observation in the data set.

(e) The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the $R^2$ and the adjusted $R^2$. Note that there are 1,236 observations in the data set.

- a:

$$\hat{bwt} = -80.41 + 0.44 \times gestation - 3.33 \times parity - 0.01 \times age \\ + 1.15 \times height + 0.05 \times weight - 8.40 \times smoke$$

- b:

  $\beta_{gestation}$ : The model predicts a $0.44$ ounce increase in the birth weight of the baby for each additional day of pregnancy, all else held constant.
  $\beta_{age}$ : The model predicts a $0.01$ ounce decrease in the birth weight of the baby for each additional year in mother's age, all else held constant.

- c:

  Parity might be correlated with one of the other variables in the model, which complicates model estimation.

- d:

$$\hat{bwt}_1 = -80.41 + 0.44 \times 284 - 3.33 \times 0 - 0.01 \times 27 + 1.15 \times 62 + 0.05 \times 100 - 8.40 \times 0 = 120.58$$

, $e_1 = 120 - 120.58 = -0.58$. The model overpredicts this baby's birth weight.

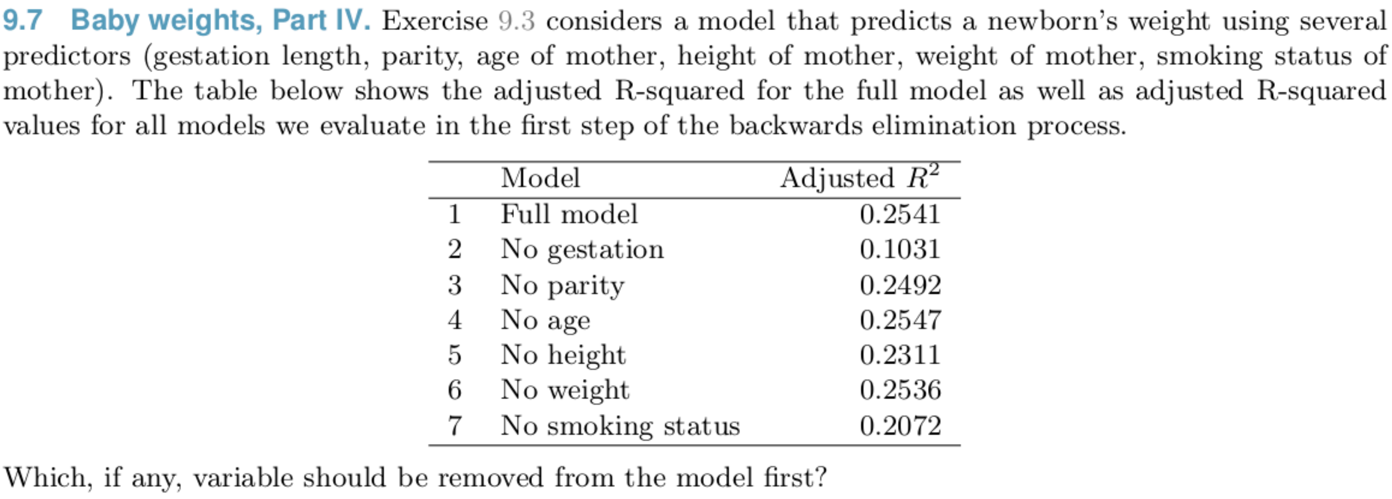
- e:

variance of the residual = $\frac{SSE}{n-1}$, variance of the data = $\frac{SST}{n-1}$

$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{249.28}{332.57} = 0.2504$

$R^2_{adj} = 1 - \frac{SSE}{SST} \times \frac{n-1}{n-p-1} = 1 - \frac{249.28}{332.57} \times \frac{1235}{1235-6} = 0.2468$

## 9.7



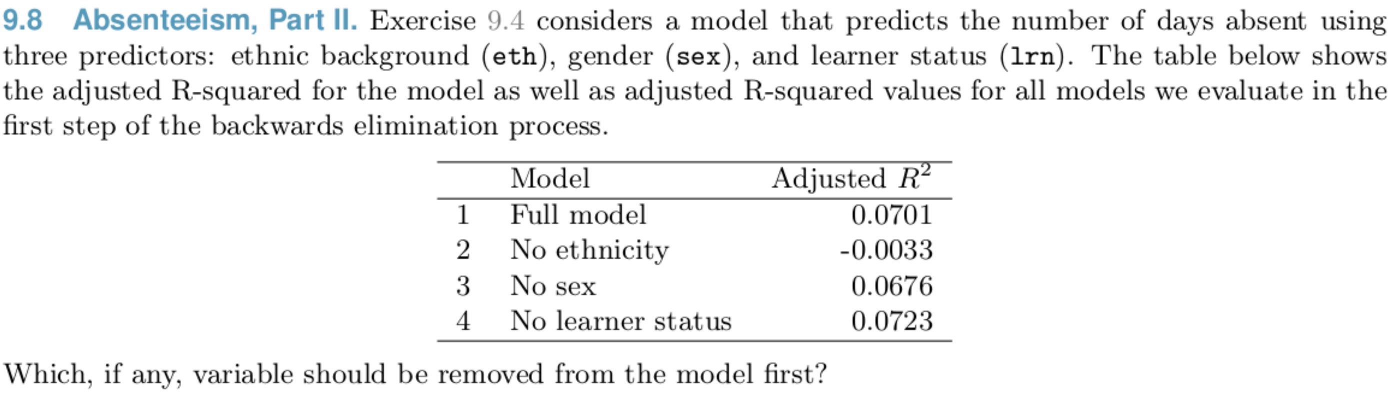
**9.7 Baby weights, Part IV.** Exercise 9.3 considers a model that predicts a newborn's weight using several predictors (gestation length, parity, age of mother, height of mother, weight of mother, smoking status of mother). The table below shows the adjusted R-squared for the full model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

| | Model | Adjusted $R^2$ |
|---|---|---|
| 1 | Full model | 0.2541 |
| 2 | No gestation | 0.1031 |
| 3 | No parity | 0.2492 |
| 4 | No age | 0.2547 |
| 5 | No height | 0.2311 |
| 6 | No weight | 0.2536 |
| 7 | No smoking status | 0.2072 |

Which, if any, variable should be removed from the model first?

Remove age, adding age as variable even cause the $R^2_{adj}$ decrease.

## 9.8



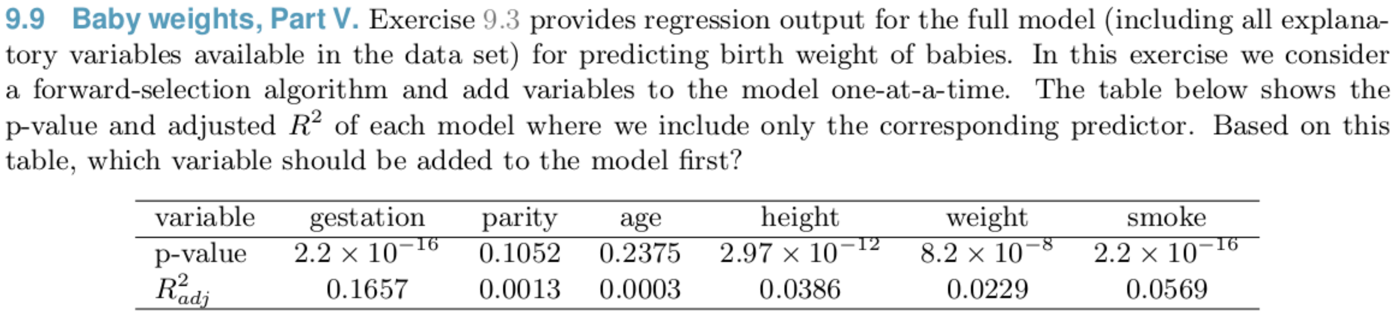
**9.8 Absenteeism, Part II.** Exercise 9.4 considers a model that predicts the number of days absent using three predictors: ethnic background (`eth`), gender (`sex`), and learner status (`lrn`). The table below shows the adjusted R-squared for the model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

| | Model | Adjusted $R^2$ |
|---|---|---|
| 1 | Full model | 0.0701 |
| 2 | No ethnicity | -0.0033 |
| 3 | No sex | 0.0676 |
| 4 | No learner status | 0.0723 |

Which, if any, variable should be removed from the model first?

Remove learner status, adding learner status as variable even cause the $R^2_{adj}$ decrease.

## 9.9

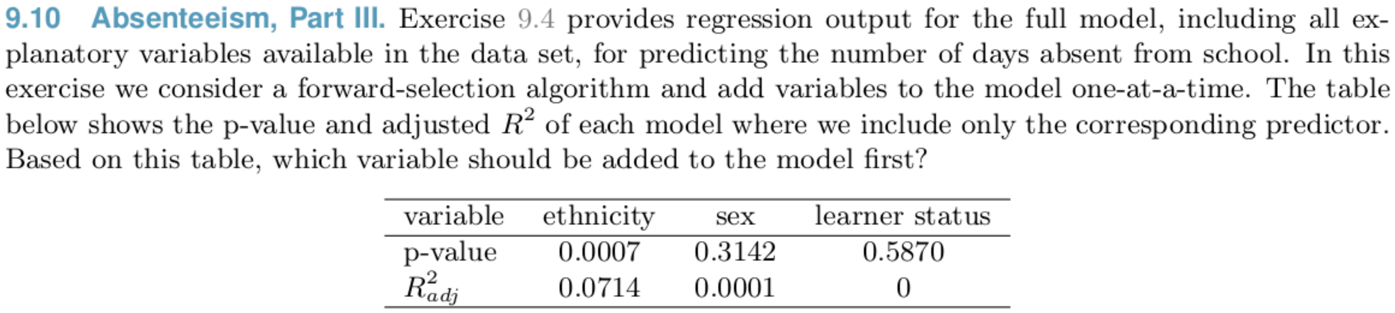


**9.9 Baby weights, Part V.** Exercise 9.3 provides regression output for the full model (including all explanatory variables available in the data set) for predicting birth weight of babies. In this exercise we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted $R^2$ of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model first?

| variable | gestation | parity | age | height | weight | smoke |
|---|---|---|---|---|---|---|
| p-value | $2.2 \times 10^{-16}$ | 0.1052 | 0.2375 | $2.97 \times 10^{-12}$ | $8.2 \times 10^{-8}$ | $2.2 \times 10^{-16}$ |
| $R^2_{adj}$ | 0.1657 | 0.0013 | 0.0003 | 0.0386 | 0.0229 | 0.0569 |

gestation, the $R^2_{adj}$ is the highest for include one variable only, and the p-value is extremely small.

## 9.10

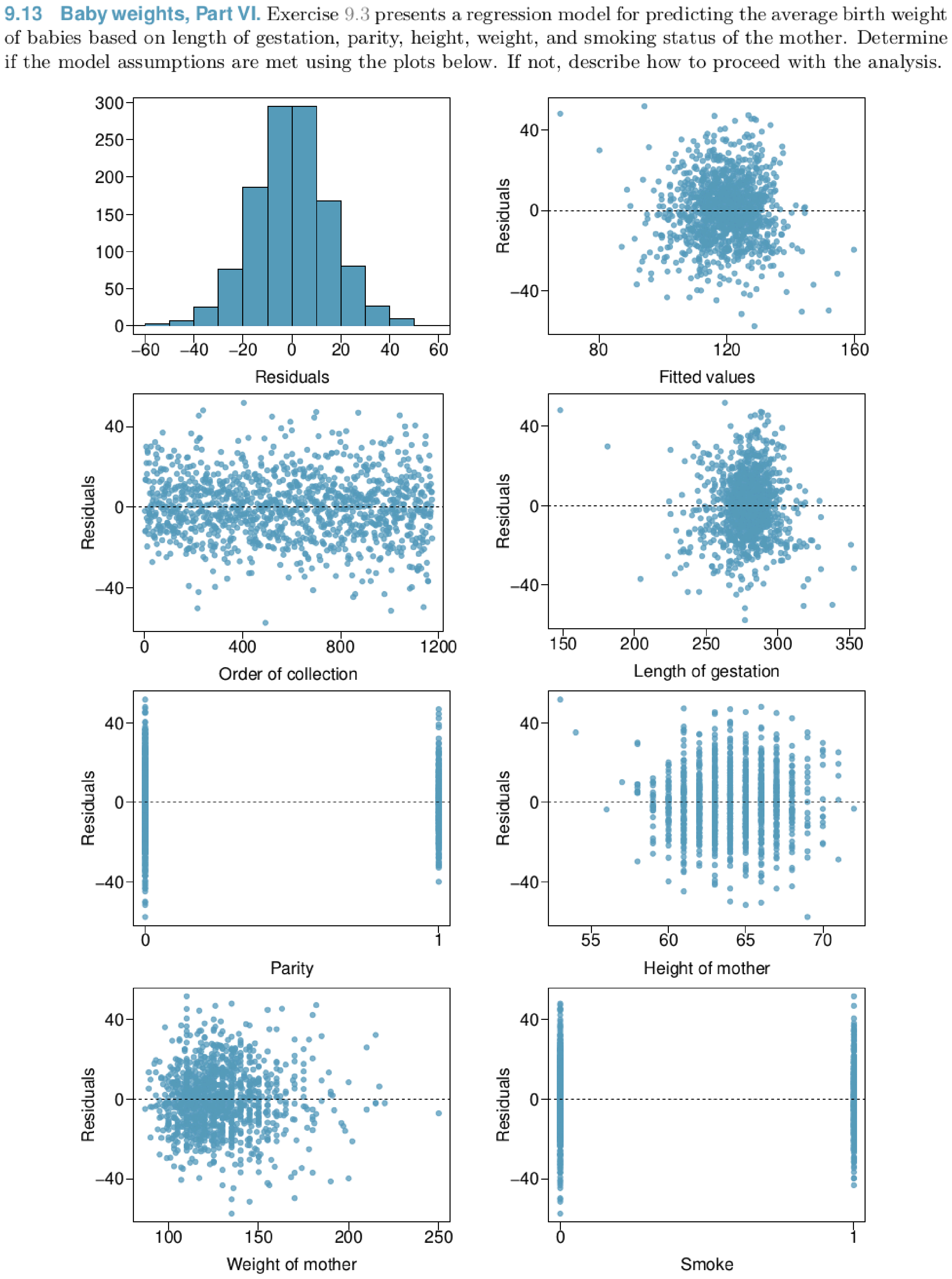


**9.10 Absenteeism, Part III.** Exercise 9.4 provides regression output for the full model, including all explanatory variables available in the data set, for predicting the number of days absent from school. In this exercise we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted $R^2$ of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model first?

| variable | ethnicity | sex | learner status |
|---|---|---|---|
| p-value | 0.0007 | 0.3142 | 0.5870 |
| $R^2_{adj}$ | 0.0714 | 0.0001 | 0 |

ethnicity, the $R^2_{adj}$ is the highest for include one variable only, and the p-value is the smallest(<0.05).

## 9.13



**9.13 Baby weights, Part VI.** Exercise 9.3 presents a regression model for predicting the average birth weight of babies based on length of gestation, parity, height, weight, and smoking status of the mother. Determine if the model assumptions are met using the plots below. If not, describe how to proceed with the analysis.

- Nearly normal residuals:
  With so many observations in the data set, we look for particularly extreme outliers in the histogram and do not see any.

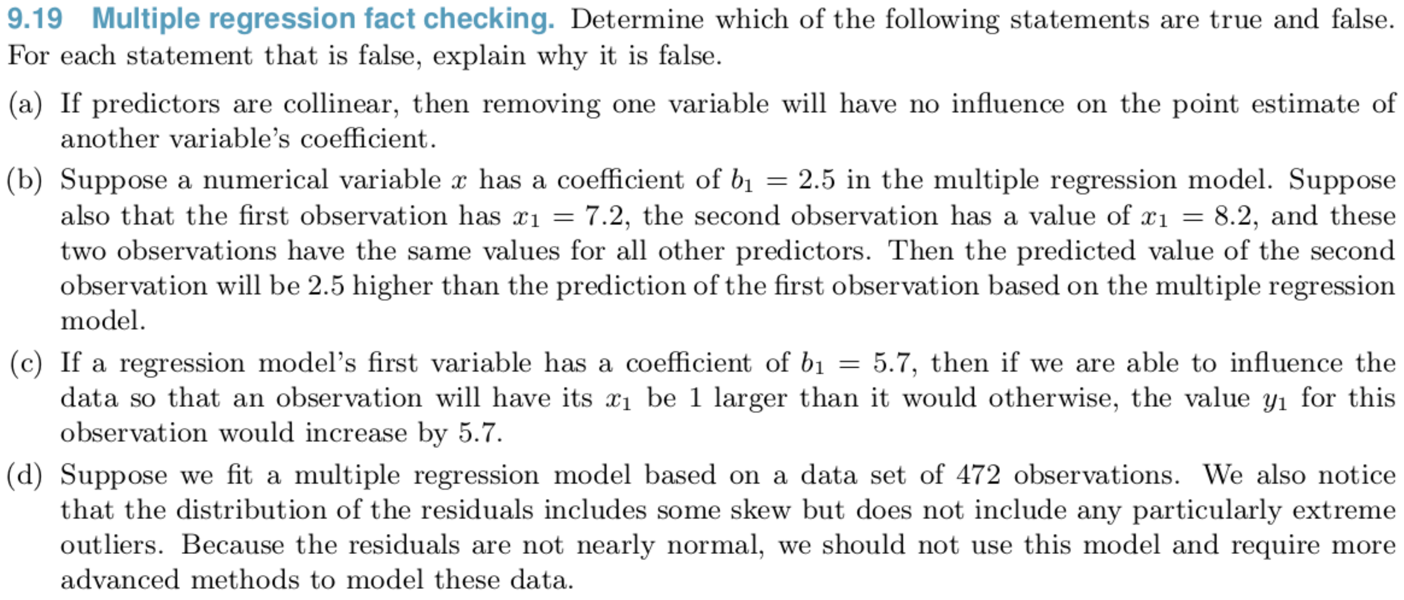
- Variability of residuals:
  The scatterplot of the residuals versus the fitted values does not show any overall structure. However, values that have very low or very high fitted values appear to also have somewhat larger outliers. In addition, the residuals do appear to have constant variability between the two parity and smoking status groups, though these items are relatively minor.

- Independent residuals:
  The scatterplot of residuals versus the order of data collection shows a random scatter, suggesting that there is no apparent structures related to the order the data were collected.

- Linear relationships between the response variable and numerical explanatory variables:
  The residuals vs. height and weight of mother are randomly distributed around 0. The residuals vs. length of gestation plot also does not show any clear or strong remaining structures, with the possible exception of
  very short or long gestations. The rest of the residuals do appear to be randomly distributed around 0.

All concerns raised here are relatively mild. There are some outliers, but there is so much data that the influence of such observations will be minor.

## 9.19



**9.19 Multiple regression fact checking.** Determine which of the following statements are true and false. For each statement that is false, explain why it is false.

(a) If predictors are collinear, then removing one variable will have no influence on the point estimate of another variable's coefficient.

(b) Suppose a numerical variable $x$ has a coefficient of $b_1 = 2.5$ in the multiple regression model. Suppose also that the first observation has $x_1 = 7.2$, the second observation has a value of $x_1 = 8.2$, and these two observations have the same values for all other predictors. Then the predicted value of the second observation will be 2.5 higher than the prediction of the first observation based on the multiple regression model.

(c) If a regression model's first variable has a coefficient of $b_1 = 5.7$, then if we are able to influence the data so that an observation will have its $x_1$ be 1 larger than it would otherwise, the value $y_1$ for this observation would increase by 5.7.

(d) Suppose we fit a multiple regression model based on a data set of 472 observations. We also notice that the distribution of the residuals includes some skew but does not include any particularly extreme outliers. Because the residuals are not nearly normal, we should not use this model and require more advanced methods to model these data.

- a: False. When predictors are collinear, it means they are correlated, and the inclusion of one variable can have a substantial influence on the point estimate (and standard error) of another.
- b: True.
- c: False. This would only be the case if the data was from an experiment and $x_1$ was one of the variables set by the researchers. (Multiple regression can be useful for forming hypotheses about causal relationships, but it offers zero guarantees.)

- d: False. We should check normality like we would for inference for a single mean: we look for particularly extreme outliers if $n \geq 30$ or for clear outliers if $n < 30$.

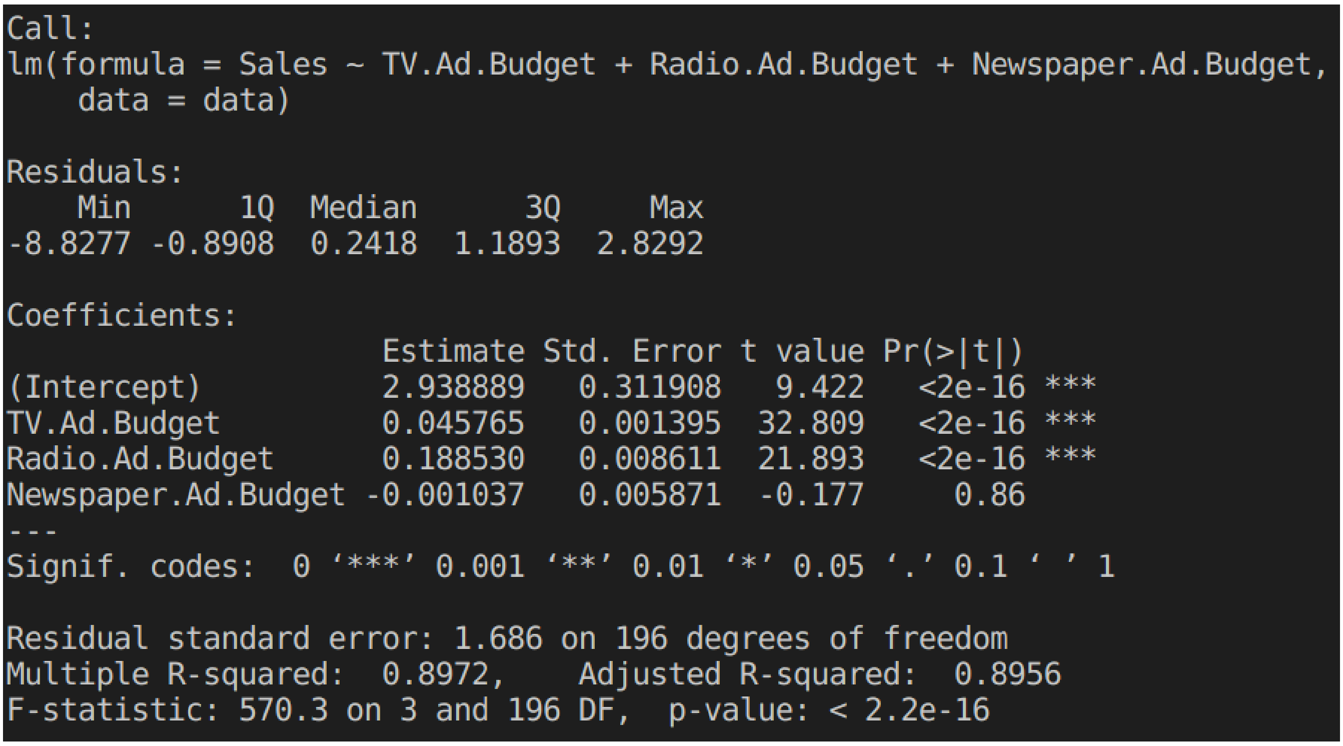
## Problem 3

### a:

- code:

```
library(readr)
library(car)

# Load the dataset
data <- read_csv('adv.csv')

# Change the column names to remove the spaces
colnames(data) <- c('TV.Ad.Budget', 'Radio.Ad.Budget', 'Newspaper.Ad.Budget', 'Sa

print(head(data))

# Perform multiple linear regression
model <- lm(Sales ~ TV.Ad.Budget + Radio.Ad.Budget + Newspaper.Ad.Budget, data =

# Display the summary of the regression model
print(summary(model))
```

- result:



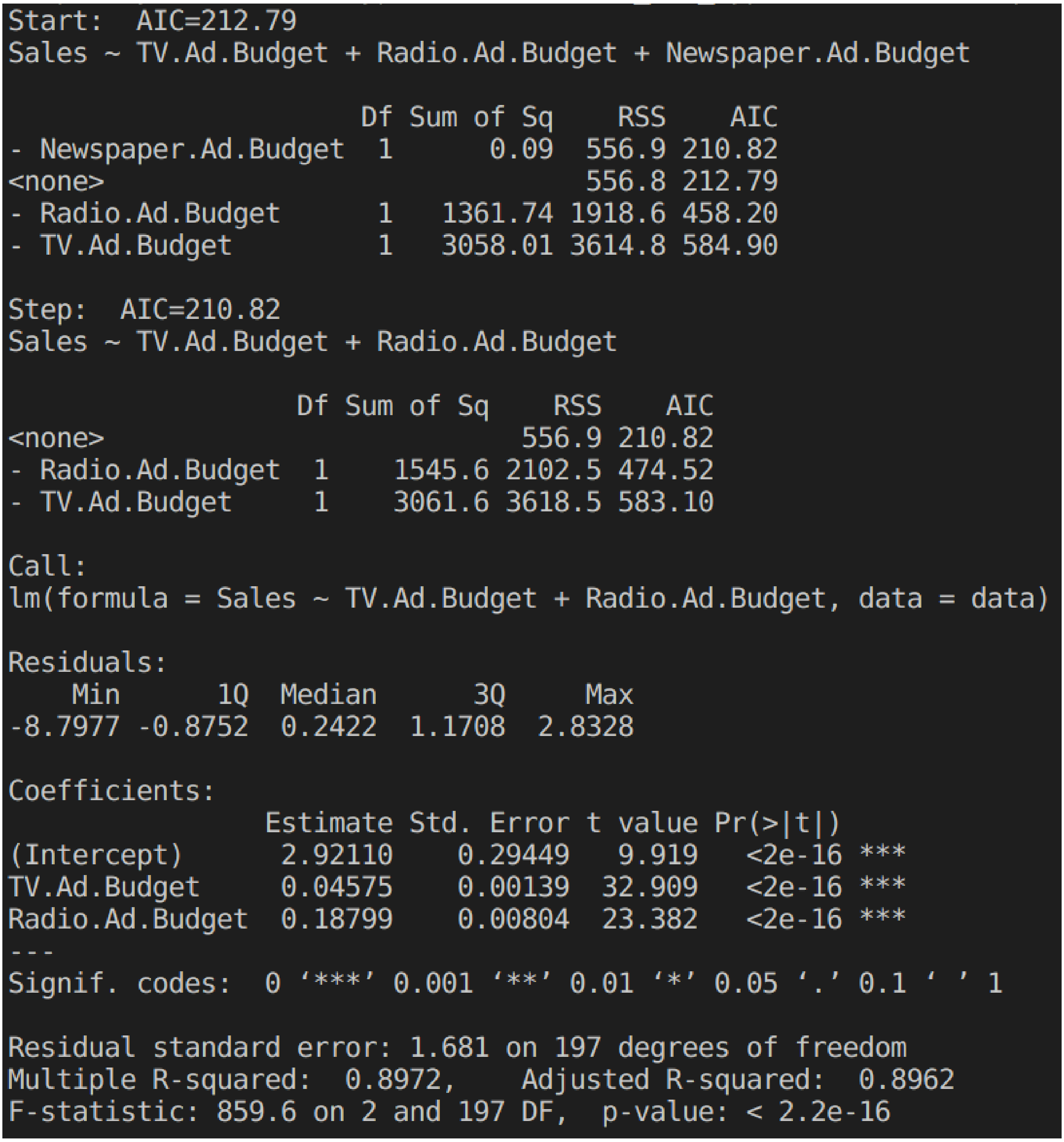
```
Call:
lm(formula = Sales ~ TV.Ad.Budget + Radio.Ad.Budget + Newspaper.Ad.Budget,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8277 -0.8908  0.2418  1.1893  2.8292 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          2.938889   0.311908   9.422   <2e-16 ***
TV.Ad.Budget         0.045765   0.001395  32.809   <2e-16 ***
Radio.Ad.Budget      0.188530   0.008611  21.893   <2e-16 ***
Newspaper.Ad.Budget -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

### b:

- code:

```
# Perform backward elimination
reduced_model <- step(model, direction = "backward")

# Display the summary of the reduced model
print(summary(reduced_model))
```

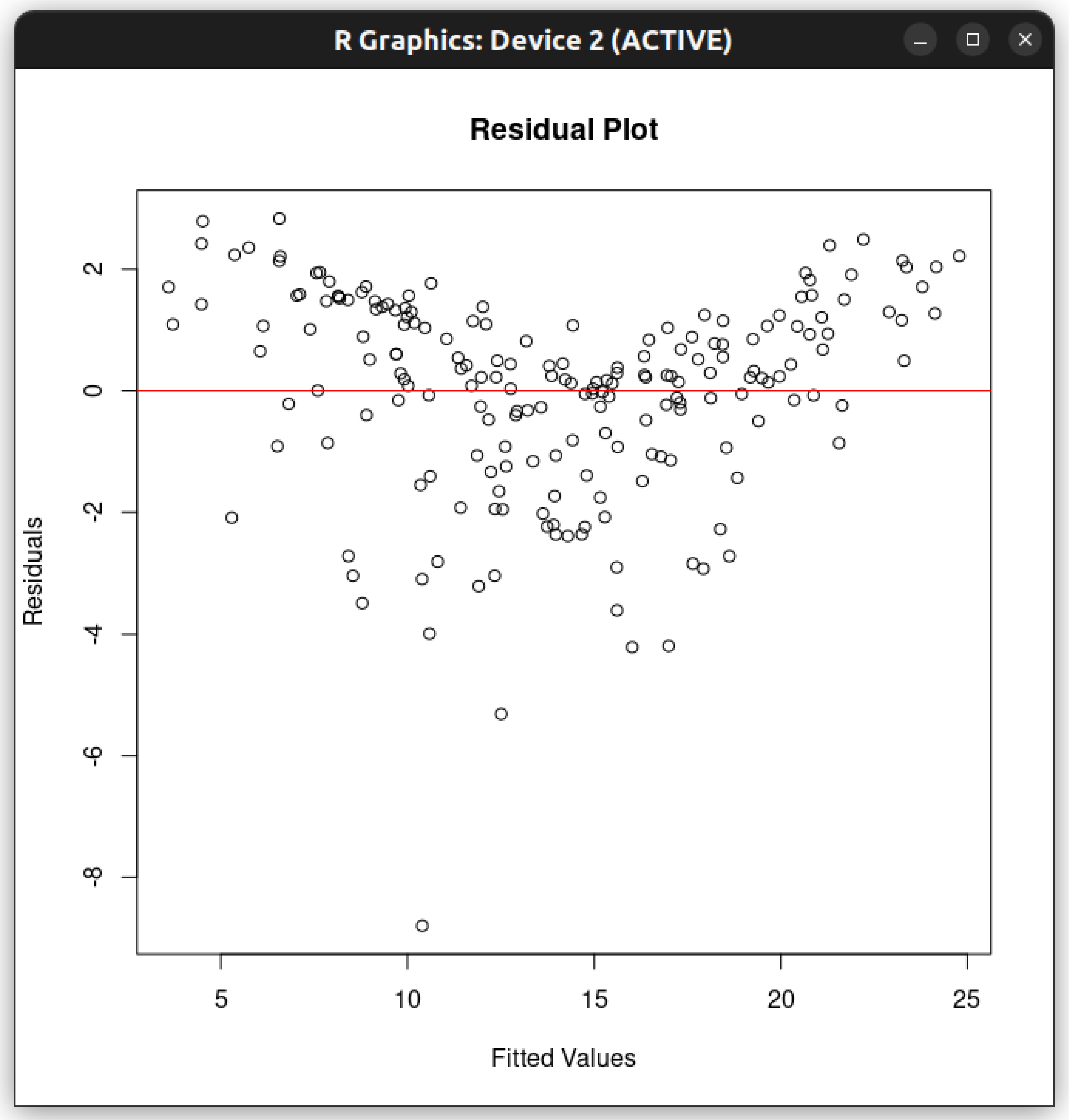
- result:

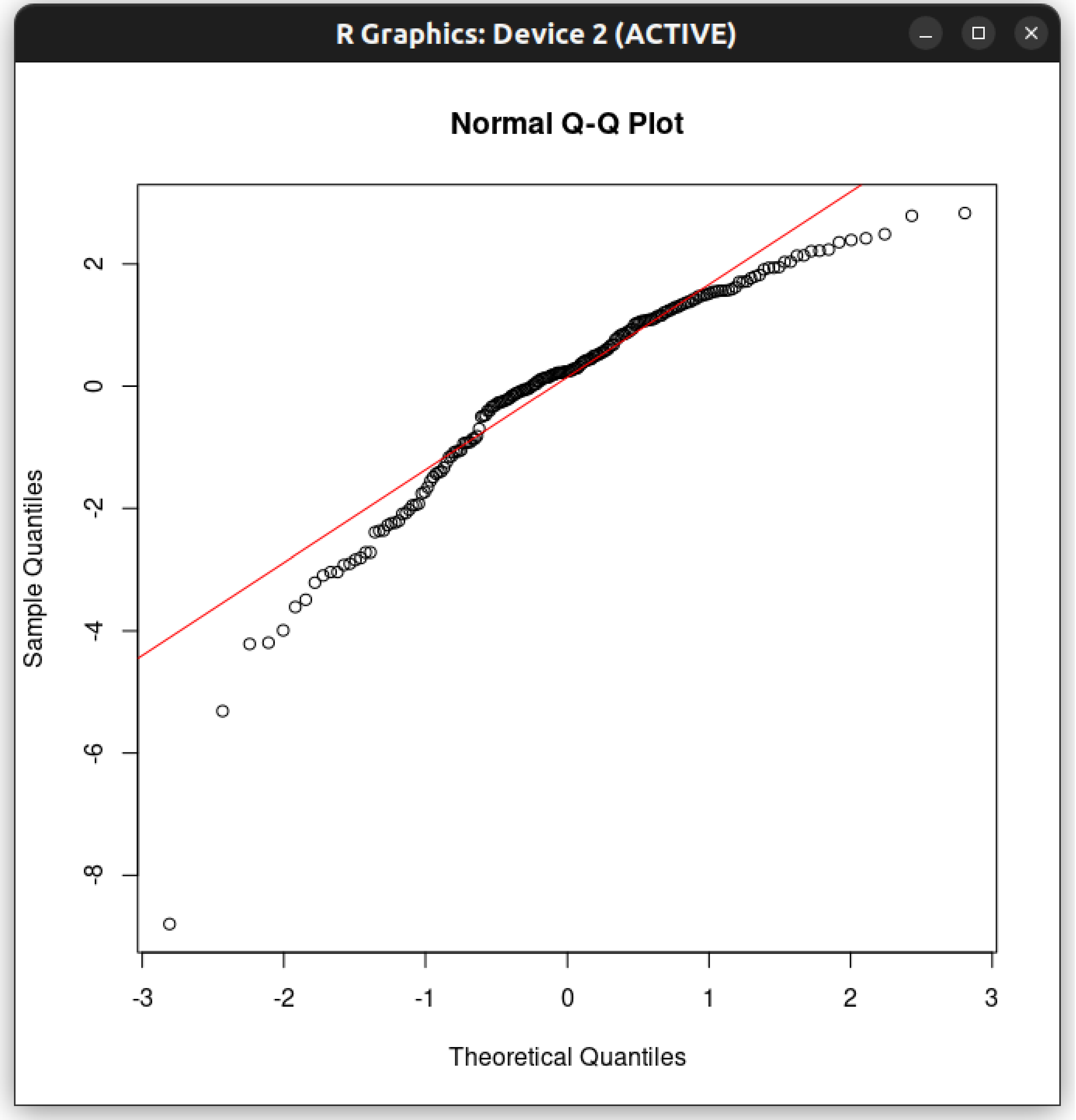


```
Start:  AIC=212.79
Sales ~ TV.Ad.Budget + Radio.Ad.Budget + Newspaper.Ad.Budget

                      Df Sum of Sq    RSS    AIC
- Newspaper.Ad.Budget  1      0.09  556.9 210.82
<none>                              556.8 212.79
- Radio.Ad.Budget      1   1361.74 1918.6 458.20
- TV.Ad.Budget         1   3058.01 3614.8 584.90

Step:  AIC=210.82
Sales ~ TV.Ad.Budget + Radio.Ad.Budget

                  Df Sum of Sq    RSS    AIC
<none>                          556.9 210.82
- Radio.Ad.Budget  1    1545.6 2102.5 474.52
- TV.Ad.Budget     1    3061.6 3618.5 583.10

Call:
lm(formula = Sales ~ TV.Ad.Budget + Radio.Ad.Budget, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.7977 -0.8752  0.2422  1.1708  2.8328 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      2.92110    0.29449   9.919   <2e-16 ***
TV.Ad.Budget     0.04575    0.00139  32.909   <2e-16 ***
Radio.Ad.Budget  0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962 
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

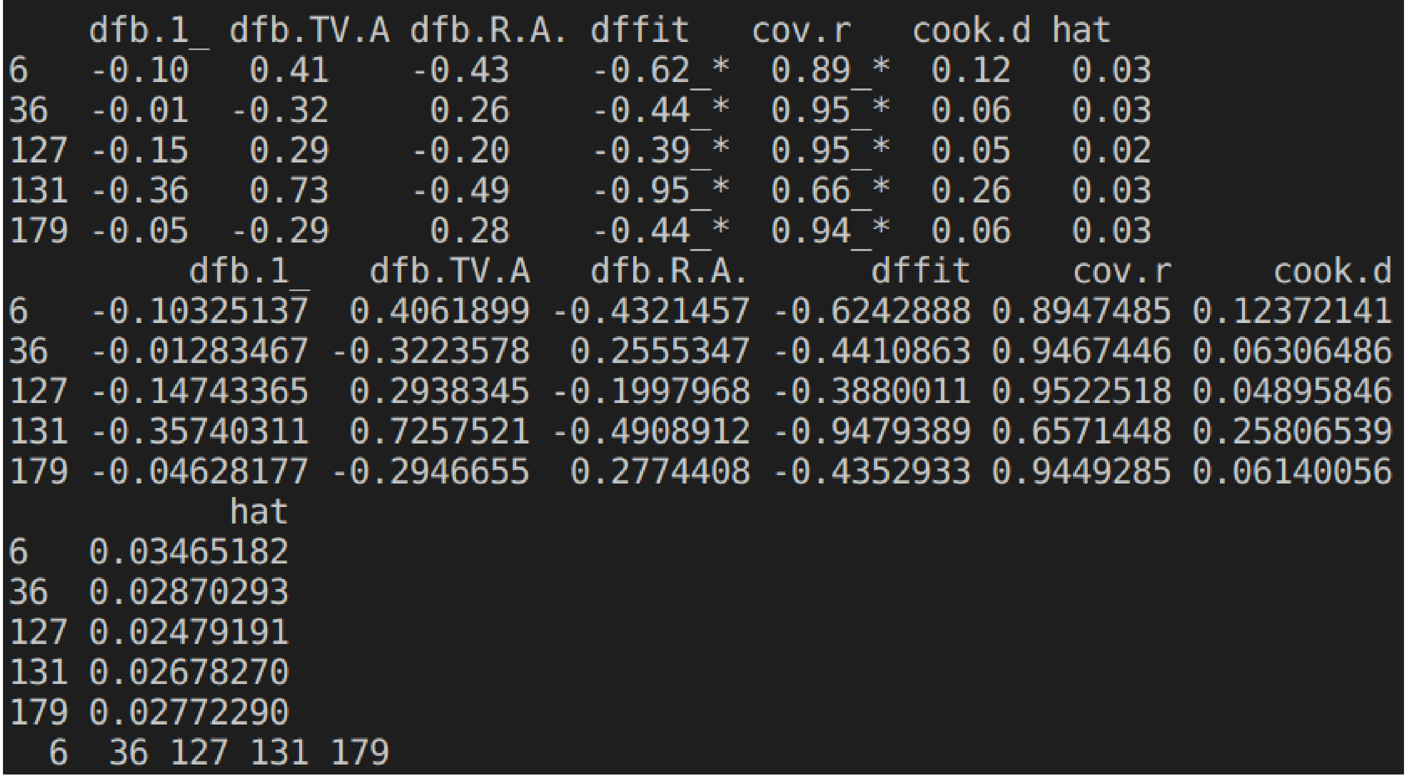
Since the p-value of Newspaper.Ad.Budget is large, we drop Newspaper.Ad.Budget(and AIC).

c:


R Graphics: Device 2 (ACTIVE)
Residual Plot
Residuals
Fitted Values
2
0
-2
-4
-6
-8
5
10
15
20
25



It seems to have visual influential observations.

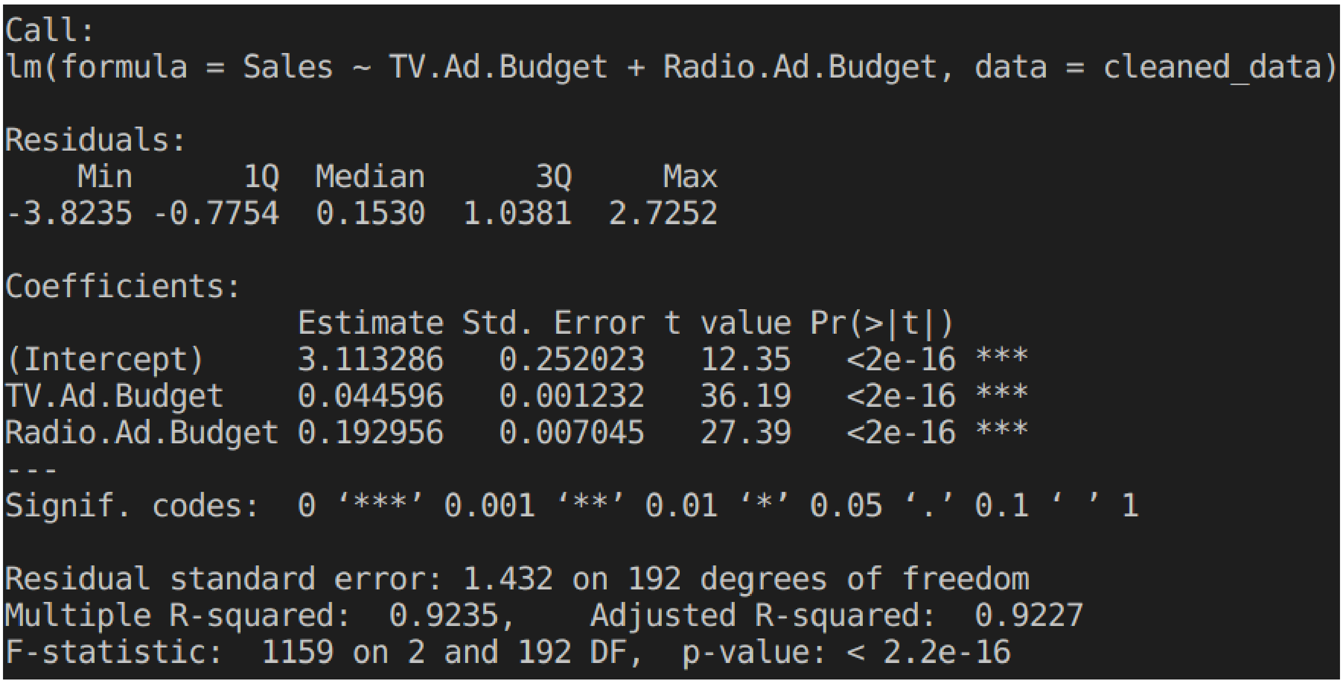


```
    dfb.1_ dfb.TV.A dfb.R.A. dffit   cov.r   cook.d hat 
6   -0.10   0.41    -0.43    -0.62_*  0.89_*  0.12   0.03
36  -0.01  -0.32     0.26    -0.44_*  0.95_*  0.06   0.03
127 -0.15   0.29    -0.20    -0.39_*  0.95_*  0.05   0.02
131 -0.36   0.73    -0.49    -0.95_*  0.66_*  0.26   0.03
179 -0.05  -0.29     0.28    -0.44_*  0.94_*  0.06   0.03
         dfb.1_    dfb.TV.A   dfb.R.A.      dffit     cov.r     cook.d
6   -0.10325137  0.4061899 -0.4321457 -0.6242888 0.8947485 0.12372141
36  -0.01283467 -0.3223578  0.2555347 -0.4410863 0.9467446 0.06306486
127 -0.14743365  0.2938345 -0.1997968 -0.3880011 0.9522518 0.04895846
131 -0.35740311  0.7257521 -0.4908912 -0.9479389 0.6571448 0.25806539
179 -0.04628177 -0.2946655  0.2774408 -0.4352933 0.9449285 0.06140056
           hat
6   0.03465182
36  0.02870293
127 0.02479191
131 0.02678270
179 0.02772290
  6  36 127 131 179
```

- code:

```
# Residual plot
plot(reduced_model$fitted.values, residuals(reduced_model), main = "Residual Plot
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red")

# Q-Q plot
qqnorm(residuals(reduced_model))
qqline(residuals(reduced_model), col = "red")

# Identify influential observations
influence <- influence.measures(reduced_model)
# print(summary(influence))

# Remove influential observations
influential_obs <- which(apply(influence$is.inf, 1, any))
print(influential_obs)
cleaned_data <- data[-influential_obs, ]
# print(head(cleaned_data))

# Refit the model
final_model <- lm(Sales ~ TV.Ad.Budget + Radio.Ad.Budget, data = cleaned_data)

# Display the summary of the final model
print(summary(final_model))
```

- result:



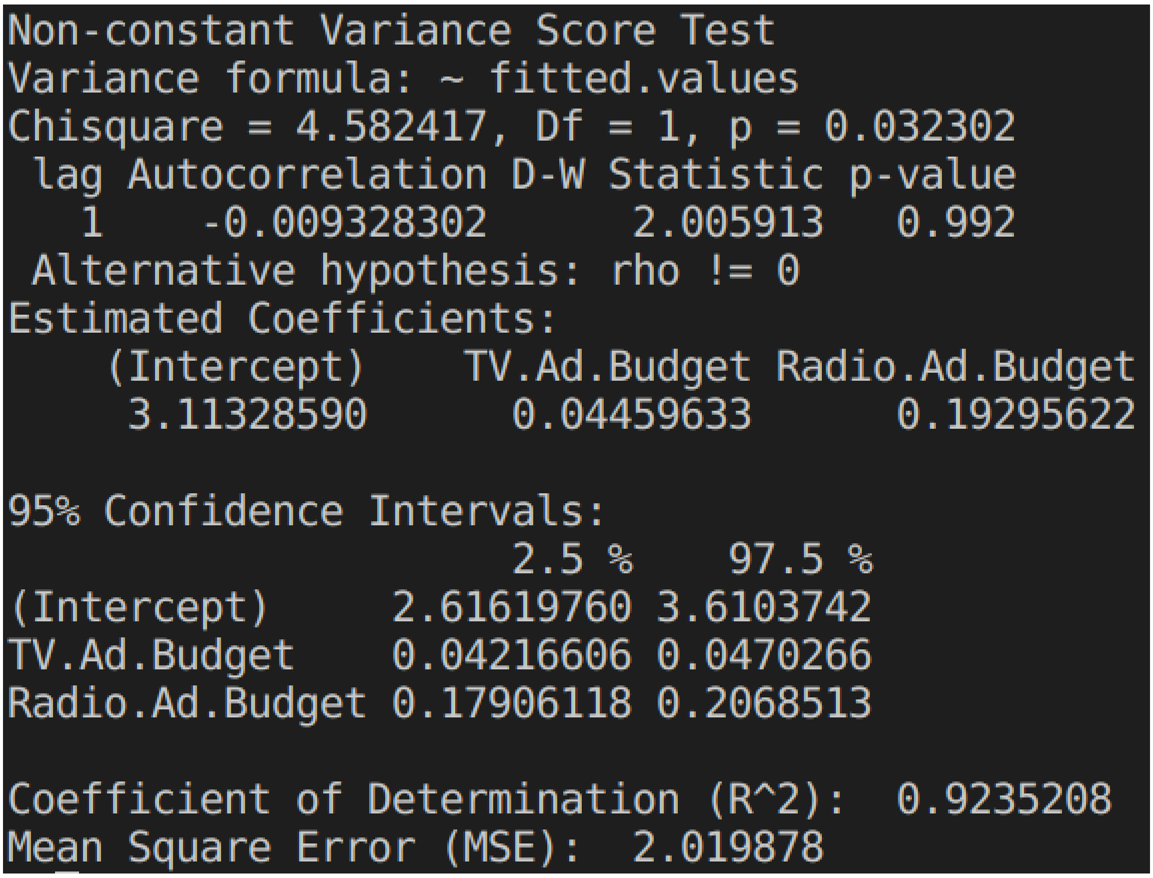
```
Call:
lm(formula = Sales ~ TV.Ad.Budget + Radio.Ad.Budget, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8235 -0.7754  0.1530  1.0381  2.7252 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.113286   0.252023   12.35   <2e-16 ***
TV.Ad.Budget    0.044596   0.001232   36.19   <2e-16 ***
Radio.Ad.Budget 0.192956   0.007045   27.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.432 on 192 degrees of freedom
Multiple R-squared:  0.9235,	Adjusted R-squared:  0.9227 
F-statistic:  1159 on 2 and 192 DF,  p-value: < 2.2e-16
```

## d:

- code:

```
# Test the assumptions of the linear regression model
## Test constant variance hypothesis
ncvTest(final_model) # reject H_0 at level 0.05
print(ncvTest(final_model))

## Test the independent of residuals
durbinWatsonTest(final_model)
print(durbinWatsonTest(final_model))

# (i) Estimated coefficients
coefficients(final_model)

# (ii) 95% confidence intervals for the coefficients
confint(final_model)

# (iii) Coefficient of determination (R^2) and Mean Square Error (MSE)
r_squared <- summary(final_model)$r.squared
mse <- mean(residuals(final_model)^2)

# Display results
cat("Estimated Coefficients:\n")
print(coefficients(final_model))
cat("\n95% Confidence Intervals:\n")
print(confint(final_model))
cat("\nCoefficient of Determination (R^2): ", r_squared, "\n")
cat("Mean Square Error (MSE): ", mse, "\n")
```

- result:



```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 4.582417, Df = 1, p = 0.032302
 lag Autocorrelation D-W Statistic p-value
   1    -0.009328302      2.005913   0.992
 Alternative hypothesis: rho != 0
Estimated Coefficients:
    (Intercept)    TV.Ad.Budget Radio.Ad.Budget
     3.11328590      0.04459633      0.19295622

95% Confidence Intervals:
                     2.5 %    97.5 %
(Intercept)     2.61619760 3.6103742
TV.Ad.Budget    0.04216606 0.0470266
Radio.Ad.Budget 0.17906118 0.2068513

Coefficient of Determination (R^2):  0.9235208
Mean Square Error (MSE):  2.019878
```