XIAMEN UNIVERSITY MALAYSIA

Course Code	:	MAT314
Course Name	:	Regression Analysis
Lecturer	:	Lim Min
Academic Session	:	2026/04
Assessment Title	:	Project 1: Simple Linear Regression
Submission Due Date	:	05/05/2026
Prepared by	:	
Date Received	:	05/05/2026

Student ID	Student Name
MAT2309449	Wang Fangyu

Feedback from Lecturer :

MARKING RUBRICS

Component Title	Report				Total Marks	100
Criteria	Score and Descriptors					Marks
	Excellent	Good	Average	Need Improvement	Poor	
Introduction	10	8	6	4	2	
Data Analysis - exploratory	10	8	6	4	2	
Data Analysis - formal test	10	8	6	4	2	
Data Analysis - validation	10	8	6	4	2	
Data Analysis - compare models	10	8	6	4	2	
Data Analysis - CIs and PIs	10	8	6	4	2	
Conclusions and recommendations	10	8	6	4	2	
R Code and R Output	10	8	6	4	2	
Use of language is appropriate, with no grammatical errors.	10	8	6	4	2	
Overall work is neat, well-organised and creative.	10	8	6	4	2	
					TOTAL	

Own Work Declaration

I acknowledge that my work will be examined for plagiarism or any other form of academic misconduct, and that the digital copy of the work may be retained for future comparisons.

I confirm that all sources cited in this work have been correctly acknowledged and that I fully understand the serious consequences of any intentional or unintentional academic misconduct.

In addition, I affirm that this submission does not contain any materials generated by AI tools, including direct copying and pasting of text or paraphrasing. This work is my original creation, and it has not been based on any work of other students (past or present), nor has it been previously submitted to any other course or institution.

Signature : 

Date : 05/05/2026

1. Introduction

The primary purpose of this project is to search the impact of different advertising expenditures on product sales using Simple Linear Regression (SLR). By analyzing historical data, the best fit for the model (SLR) is determined among the three potential regressors that can potentially influence sales: TV, radio, and newspaper, and an SLR model of such a regressor is constructed to forecast future sales based on specific budgets.

The dataset contains 200 observations, each representing the annual sales (in thousands of units) from a different market, which is the response variable, and their corresponding three types of annual advertising expenditures, precisely, TV, radio, and newspaper advertising expenditures (in thousands of dollars), which are the regressor variables. To verify the fit of the model, the dataset was split randomly into two distinct subsets. The estimation dataset, containing 160 observations (80% of the total data), was used to fit the model. The remaining 40 observations constitute the prediction dataset, which was used to verify the accuracy of the fitted model.

To achieve the project purpose, the student ID was first used as the random seed in R to split the original dataset into two sets. Next, scatter diagrams were constructed for each regressor versus sales of the estimation set to visualize the underlying potential linear relationships. Then, three sample correlation coefficients were computed by R for each regressor versus sales to get a general understanding of the correlation. After that, three simple linear regression models, one for each pair: TV versus Sales, Radio versus Sales, Newspaper versus Sales with Sales as the response variable, were fitted using R based on the estimation set. Then, three sets of the predicted sales for values of each corresponding regressor in the prediction data were fitted. Furthermore, three scatter diagrams for sales from the prediction data versus predicted sales were constructed by R to visualize the general accuracy of the fitted models. Moreover, the best model was selected based on (R^2), residual standard error, and validation results obtained by R. Eventually, two 95% confidence intervals of two different amounts spent on advertising for the mean sales were computed using the selected best model; in addition, two 95% prediction intervals for the sales next year of two different budgets on advertising were also computed using the selected model.

2. Data Analysis

To begin the analysis, the student ID was first used as the random seed in R to split the original dataset into two sets: estimation dataset (80%, used to build the regression models) and prediction dataset (20%, used to validate the models).

2.1. **Scatter Diagrams & Sample Correlation Coefficients.**

The estimation dataset is used in the following analysis. Three scatter diagrams and the corresponding sample correlation coefficient, one for each pair: TV versus Sales, Radio versus Sales, Newspaper versus Sales are produced and displayed in Figure 1, Figure 2, Figure 3 and Table 1.

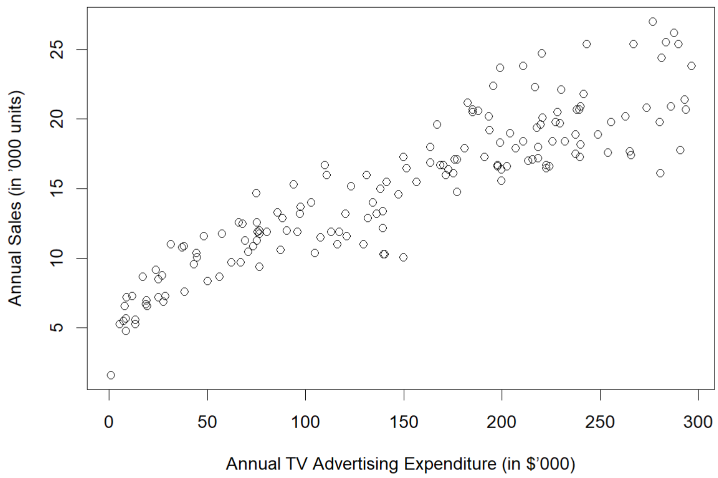


Figure 1. Scatter Diagram of TV Advertising Expenditures and Sales

Figure 1 displays that there is a clear potential linear relationship between TV advertising expenditures and sales.

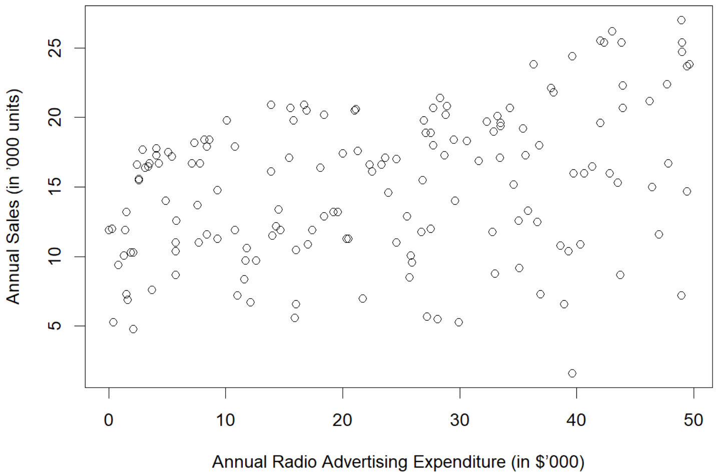


Figure 2. Scatter Diagram of Radio Advertising Expenditures and Sales

Figure 2 displays a weak evidence that exist a potential linear relationship between Radio advertising expenditures and sales.

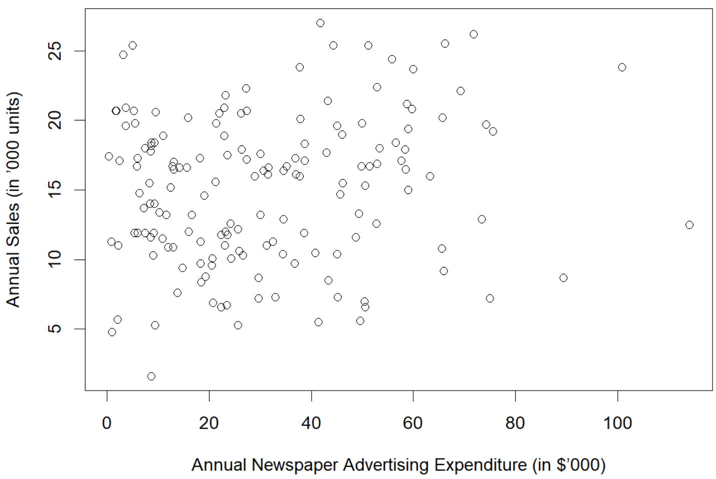


Figure 3. Scatter Diagram of Newspaper Advertising Expenditures and Sales

Figure 3 displays that there probably not exist a linear relationship between Newspaper advertising expenditures and sales.

	Sample Correlation Coefficients
TV	0.9050833
Radio	0.3551465
Newspaper	0.153969

Table 1. Sample Correlation Coefficients of Advertising Expenditures and Sales

Above are the sample correlation coefficients for the three groups computed using R, which were 0.9050833, 0.3551465, and 0.153969, respectively. These results were consistent with our observations from the scatter diagram. Since the sample correlation coefficients values were not approximately equal to 0, linear regression fitting is decided to perform on the three groups of estimation data.

2.2. Model Fitting & Analysis.

Based on the exploratory study, Sales (y) and TV Advertising Expenditures, Radio Advertising Expenditures, Newspaper Advertising Expenditures (x) are fitted to the following model based on the estimation dataset

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{1}$$

A summary of the output for TV-Sales model is shown in Table 2.

	Estimate	Std. Error	t	P-value
Intercept	6.805068	0.356732	19.08	$< 2 \times 10^{-16}$
TV	0.055836	0.002087	26.75	$< 2 \times 10^{-16}$

$$R_e = 2.254 \text{ on df} = 158$$
$$R^2 = 0.8192$$
$$F_{1,158} = 715.8,\ p\text{-value: } < 2.2 \times 10^{-16}$$

Table 2. Summary Output for TV-Sales Model

The fitted function for the TV-Sales model is

$$\hat{y} = 6.805068 + 0.055836x. \tag{2}$$

It can be concluded that when the TV advertising expenditure is 0, the estimated sales are 6.8051 thousand unit and as the TV advertising expenditures increase by 1 thousand dollars, the sales will increase by 0.0558 thousand units.

$F_{1,158} = 715.8 >> 1$ with p-value: $< 2.2 \times 10^{-16}$ suggests that there is enough evidence that TV advertising expenditure has a significant linear effect on sales.

R^2, the coefficient of determination, interpreted as the proportion of variation in y that is explained by the regressor x, is 0.8192, which implies that 81.92% of the variation in sales can be explained by the TV advertising expenditures.

Therefore, these results indicate a strong linear relationship between the two variables, suggesting that the model fits the data very well.

A summary of the output for Radio-Sales model is shown in Table 3.

	Estimate	**Std. Error**	***t***	***P*-value**
Intercept	12.1526	0.7262	16.734	$< 2 \times 10^{-16}$
Radio	0.1270	0.0266	4.775	4.06×10^{-6}

$$R_e = 4.956 \text{ on df} = 158$$
$$R^2 = 0.1261$$
$$F_{1,158} = 22.8,\ p\text{-value: } 4.062 \times 10^{-6}$$

TABLE 3. Summary Output for Radio-Sales Model

The fitted function for the Radio-Sales model is

$$\hat{y} = 12.1526 + 0.1270x. \tag{3}$$

It can be concluded that when the Radio advertising expenditure is 0, the estimated sales are 12.1526 thousand unit and as the Radio advertising expenditures increase by 1 thousand dollars, the sales will increase by 0.1270 thousand units.

$F_{1,158} = 22.8 >> 1$ with p-value: 4.062×10^{-6} suggests that there is enough evidence that Radio advertising expenditure has a significant linear effect on sales.

R^2, the coefficient of determination, interpreted as the proportion of variation in y that is explained by the regressor x, is 0.1261, which implies that 12.61% of the variation in sales can be explained by the Radio advertising expenditures.

Therefore, these results indicate a weak linear relationship between the two variables, suggesting that the model does not fit the data very well.

A summary of the output for Newspaper-Sales model is shown in Table 4.

	Estimate	**Std. Error**	t	P**-value**
Intercept	13.93566	0.71300	19.545	$< 2 \times 10^{-16}$
Newspaper	0.03680	0.01879	1.959	0.0519

$$R_e = 5.238 \text{ on df} = 158$$
$$R^2 = 0.02371$$
$$F_{1,158} = 3.837,\ p\text{-value: } 0.05191$$

TABLE 4. Summary Output for Newspaper-Sales Model

The fitted function for the Newspaper-Sales model is

$$\hat{y} = 13.93566 + 0.03680x. \tag{4}$$

It can be concluded that when the Newspaper advertising expenditure is 0, the estimated sales are 13.93566 thousand unit and as the Newspaper advertising expenditures increase by 1 thousand dollars, the sales will increase by 0.03680 thousand units.

$F_{1,158} = 3.837 > 1$ with p-value: $0.05191 > 0.05$ suggests that there is not enough evidence that Newspaper advertising expenditure has a significant linear effect on sales.

R^2, the coefficient of determination, interpreted as the proportion of variation in y that is explained by the regressor x, is 0.02371, which implies that only 2.371% of the variation in sales can be explained by the Newspaper advertising expenditures.
Therefore, these results indicate a poor linear relationship between the two variables, indicates a poor fit.

2.3. **Model Validation.**

The prediction dataset is now used to validate the models. Predicted sales are obtained by substituting the specific advertising expenditure values from the prediction dataset into the respective fitted functions fitted in the 2.2.

Consequently, three scatter diagrams for sales from the prediction data versus predicted sales were constructed by R to visualize the accuracy of the fitted models. If a fitted model possesses high predictive accuracy, the data points will cluster tightly around the 45° reference line where observed sales equal predicted sales ($y = \hat{y}$). The three scatter diagrams produced for validation are displayed in Figures 4, 5, and 6.

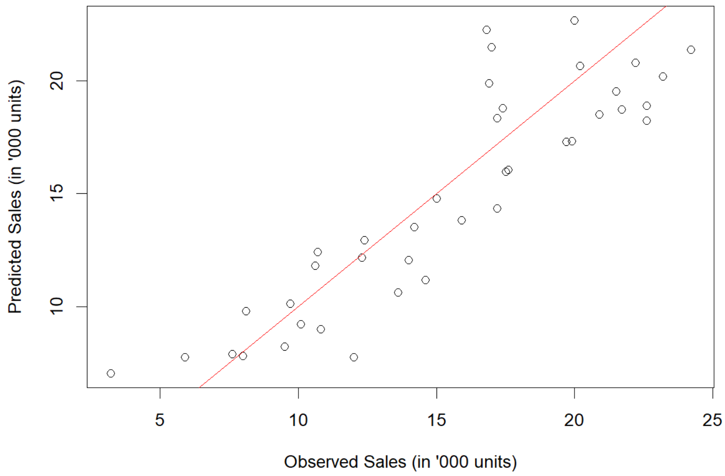


FIGURE 4. Observed Sales versus Predicted Sales for TV Model

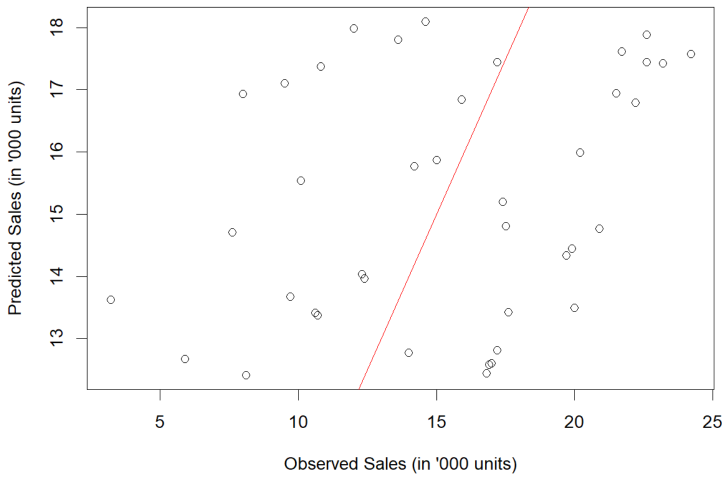


FIGURE 5. Observed Sales versus Predicted Sales for Radio Model

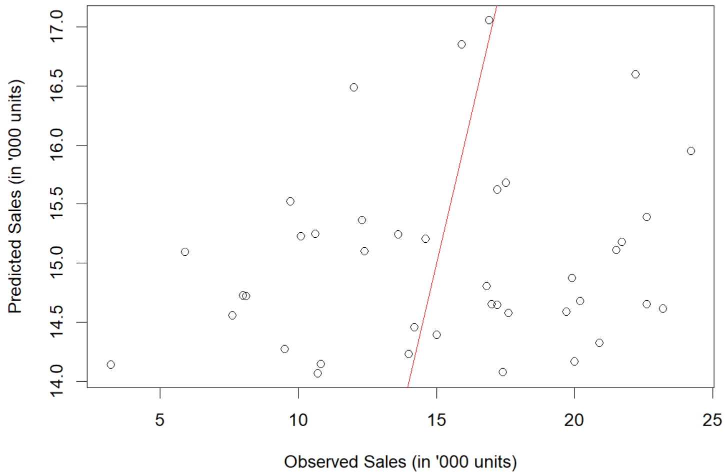


FIGURE 6. Observed Sales versus Predicted Sales for Newspaper Model

As displayed in Figure 4, the data points for the TV-Sales model align closely with the 45° reference line, indicating a high degree of predictive accuracy and implies that the model fits the new data well. In contrast, the validation plots for the Radio model (Figure 5) and the Newspaper model (Figure 6) exhibit significant dispersion away from the reference line, indicating poor degree of predictive accuracy and implies that the model does not fit the new data well, validating the previous conclusion in 2.2.

2.4. **Model Comparison and Selection.**

To select the best model among the three for predicting sales. Three properties of the models have been chosen to do the selection: the coefficient of determination (R^2), the residual standard error(R_e), and the validation results from the prediction dataset. A summary of these properties for all three fitted models is displayed in Table 5.

Model	$\boldsymbol{R^2}$	$\boldsymbol{R_e}$	**Validation Result**
TV-Sales	0.8192	2.254	Points closely align with the 45° line
Radio-Sales	0.1261	4.956	Points exhibit significant dispersion
Newspaper-Sales	0.0237	5.238	Points exhibit significant dispersion

TABLE 5. Comparison of Properties of Three Fitted Models

Based on the summary in Table 5, the TV-Sales model clearly fits better than the other two models across all three properties. It possesses the highest R^2 value (0.8192), indicating that it explains the greatest proportion of the variation in sales. Furthermore, it yields the lowest residual standard error (2.254), which implies that its predictions have the smallest average deviation from the actual observed values. Finally, it is the only one that its data points in validation plot cluster tightly around the 45° reference line.

Therefore, the TV-Sales model is selected as the best model for predicting sales. The adopted model is the simple linear regression model for TV advertising expenditures, and its fitted function is stated as follows:

$$\hat{y} = 6.8051 + 0.0558x. \tag{5}$$

2.5. **Estimation.**

Table 6 shows the 95% confidence confidence intervals and prediction intervals calculated from the adopted model based on the estimation dataset.

TV Advertising(dollar)	95% CI(unit) & width	95% PI(unit) & width
$\$151,070$	$[14,888, 15,592]\&704$	$[10,774, 19,707]\&8933$
$\$221,380$	$[18,702, 19,630]\&928$	$[14,689, 23,643]\&8954$

Table 6. 95% Confidence and Prediction Intervals for Mean and Future Sales

With 95% confidence, the mean sales is estimated to fall between $14,888$ and $15,592$ units when $\$151,070$ is spent on TV advertising and estimated to fall between $18,702$ and $19,630$ units when $\$221,380$ is spent on TV advertising. Under the same confidence level, the prediction sales is estimated to fall between $10,774$ and $19,707$ units when $\$151,070$ is the budget on TV advertising and estimated to fall between $14,689$ and $23,643$ units when $\$221,380$ is the budget on TV advertising.

By comparing the interval widths. Firstly, the prediction intervals are much wider than their corresponding confidence intervals. This difference occurs because besides the uncertainty involved in estimating the population mean, prediction interval must account for the random variation of individual future observations (ϵ). Secondly, the intervals associated with the $\$221,380$ budget are wider than those for the $\$151,070$ budget. This can be explained by the statistical principle: as the value of the predictor variable moves further away from the sample mean, the standard error of the estimate increases, therefore increasing the uncertainty of the prediction.

3. Conclusions and Recommendations

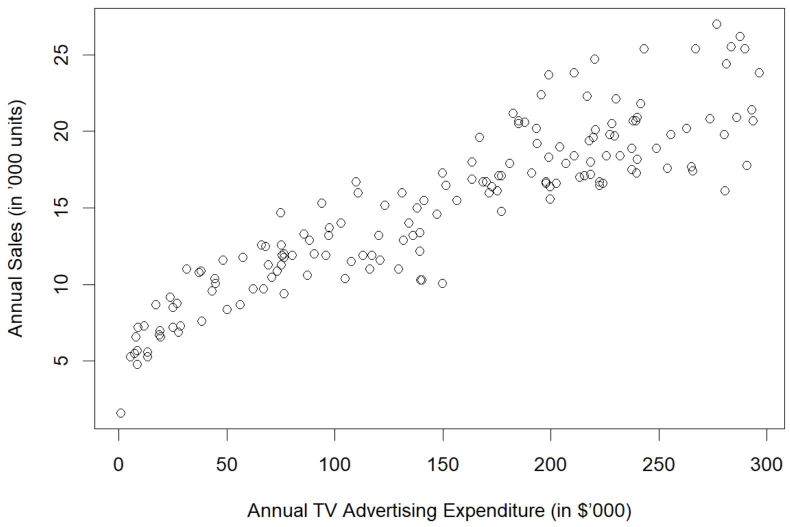
3.1. **Summary & Strengths and Weaknesses.** A simple linear regression was calculated to predict sales based on TV advertising expenditure. A significant regression equation was found ($t(158) = 26.75, p < 2\times10^{-16}$), with $R^2 = 0.8192$. The predicted sales(thousand unit) of the product is equal to 6.8051 + 0.0558 (TV advertising expenditure) when the TV advertising expenditure is measured in thousands of dollars. As the TV advertising expenditure increases by one thousand dollars, its sales are estimated to increase by 0.0558 (thousand unit). The primary strengths of the Best Model is its high explanatory power, with an R^2 of 0.8192, which is significantly higher than the R^2 of Radio and Newspaper model. The value of R^2 of TV model indicate that TV advertising expenditure can explain approximately 81.92% of the variation in sales. Besides, the TV model has a lower residual standard error (2.254), making it more reliable for budget allocations. Furthermore, the validation scatter plot confirmed its accuracy, as the plot points lies aligned to the reference line. There are also several weaknesses of the Best Model. First, it assumes that the effects of TV advertising, Radio advertising, and Newspaper advertising are independent, the potential correlation between these three predictors was ignored. Second, the prediction intervals is too wide to make it to be practical. Finally, since it is a simple linear regression, consequently, as the predictor keeps increasing, the response variable will keeps increasing unlimitly, which is arbitrary to the real life.

3.2. **Recommendations.** To overcome the weaknesses mention above, there are few recommendations:

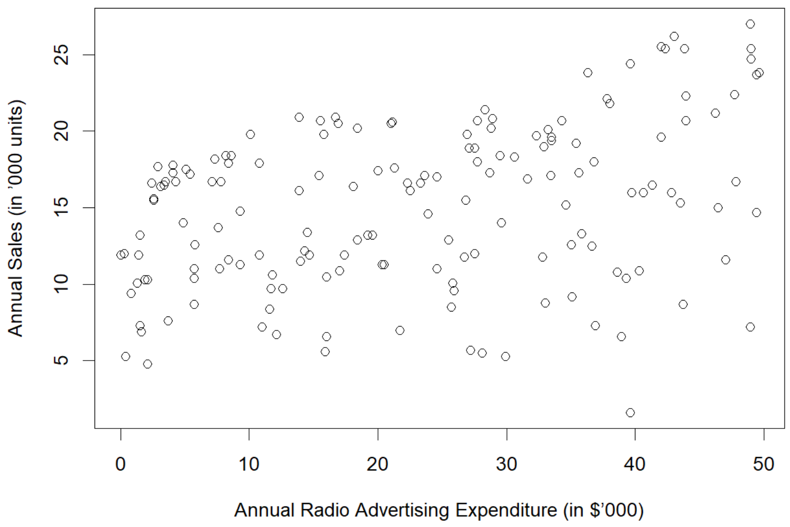
(1) **Correlation Consideration**: Improve the model selection to take into account the ability to fit data from TV, radio, and newspaper simultaneously.
(2) **Data Improvement**: Collect more and more accurate data to improve the fitting effect, therefore reducing the problem of too wide prediction interval.
(3) **Model Selection**: Consider a non-linear model to address the impractical problem of the unlimited growing response variable. Maybe logarithm function.

4. R code and R output

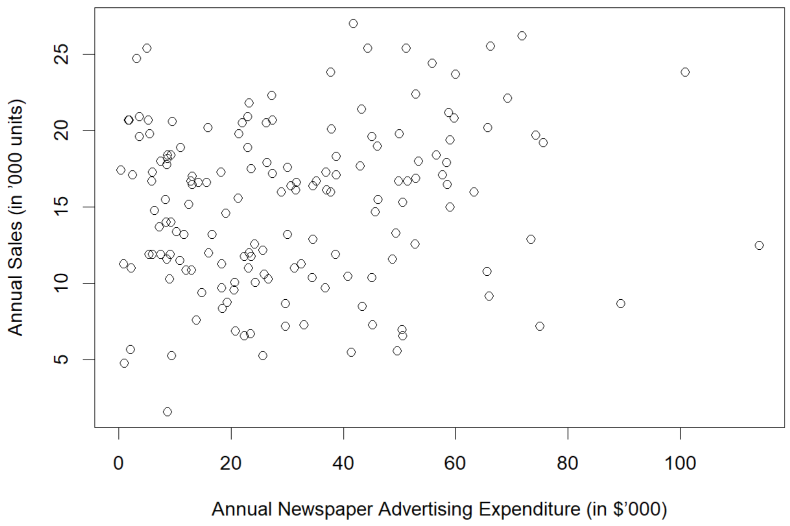
```
AdData <- read.csv("advertising.csv", header = TRUE)
> set.seed(2309449)
> dt <- sort(sample(nrow(AdData), nrow(AdData)*.8))
> AdData_e <- AdData[ dt, ]
> AdData_p <- AdData[-dt, ]
> plot(AdData_e$TV, AdData_e$Sales,
+      xlab = "Annual TV Advertising Expenditure (in $'000)",
+      ylab = "Annual Sales (in '000 units)")
>
```




```
> plot(AdData_e$Radio, AdData_e$Sales,
+      xlab = "Annual Radio Advertising Expenditure (in $'000)",
+      ylab = "Annual Sales (in '000 units)")
>
```



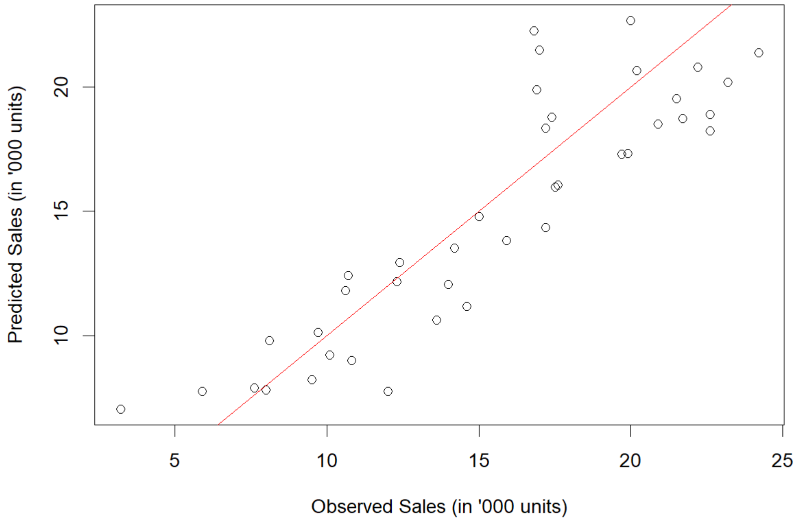
```
> plot(AdData_e$Newspaper, AdData_e$Sales,
+      xlab = "Annual Newspaper Advertising Expenditure (in $'000)",
+      ylab = "Annual Sales (in '000 units)")
>
```



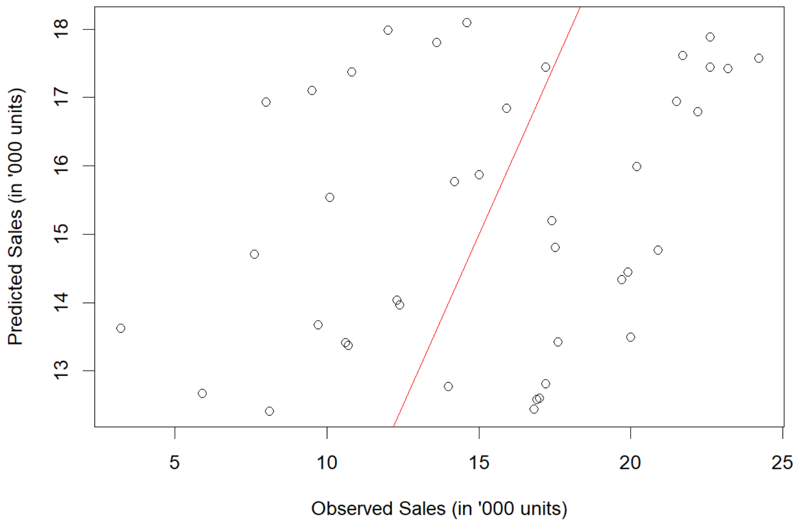

```
    > cor(AdData_e$TV, AdData_e$Sales)
[1] 0.9050833
> cor(AdData_e$Radio, AdData_e$Sales)
[1] 0.3551465
> cor(AdData_e$Newspaper, AdData_e$Sales)
[1] 0.153969
> model_tv <- lm(Sales ~ TV, data = AdData_e)
> model_radio <- lm(Sales ~ Radio, data = AdData_e)
> model_news <- lm(Sales ~ Newspaper, data = AdData_e)
> summary(model_tv)

Call:
lm(formula = Sales ~ TV, data = AdData_e)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3782 -1.4427 -0.0428  1.2896  5.7892

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.805068   0.356732   19.08   <2e-16 ***
TV          0.055836   0.002087   26.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.254 on 158 degrees of freedom
Multiple R-squared:  0.8192, Adjusted R-squared:  0.818
F-statistic: 715.8 on 1 and 158 DF,  p-value: < 2.2e-16

> summary(model_radio)
```

```
Call:
lm(formula = Sales ~ Radio, data = AdData_e)

Residuals:
     Min       1Q   Median       3Q      Max
-15.5836  -3.4380   0.6643   4.1537   8.6348

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1526     0.7262  16.734  < 2e-16 ***
Radio         0.1270     0.0266   4.775 4.06e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.956 on 158 degrees of freedom
Multiple R-squared:  0.1261, Adjusted R-squared:  0.1206
F-statistic:  22.8 on 1 and 158 DF,  p-value: 4.062e-06

> summary(model_news)

Call:
lm(formula = Sales ~ Newspaper, data = AdData_e)

Residuals:
     Min       1Q   Median       3Q      Max
-12.6558  -3.7972   0.8054   3.8057  11.5262

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.93566    0.71300  19.545   <2e-16 ***
Newspaper    0.03680    0.01879   1.959   0.0519 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.238 on 158 degrees of freedom
Multiple R-squared:  0.02371, Adjusted R-squared:  0.01753
F-statistic: 3.837 on 1 and 158 DF,  p-value: 0.05191
```

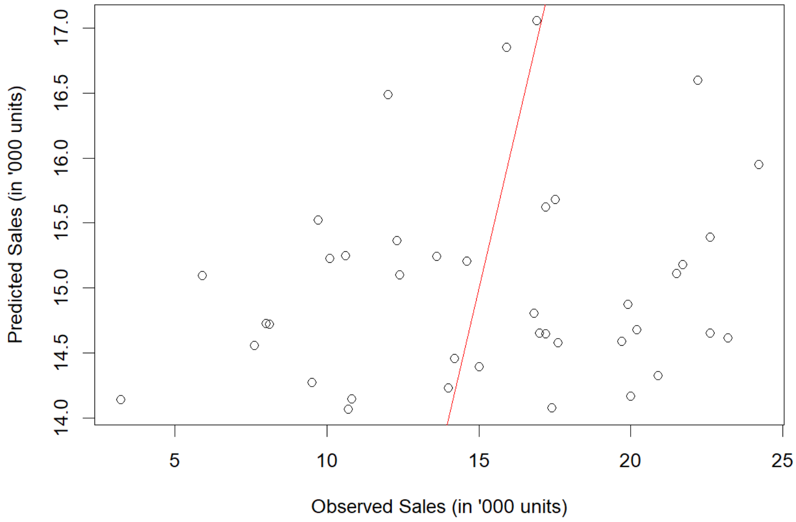
```
> pred_TV_sales <- predict(model_tv, newdata = AdData_p)
> plot(AdData_p$Sales, pred_TV_sales,
+      xlab = "Observed Sales (in '000 units)",
+      ylab = "Predicted Sales (in '000 units)")
> abline(a = 0, b = 1, col = "red")
>
```




```
> pred_radio_sales <- predict(model_radio, newdata = AdData_p)
> plot(AdData_p$Sales, pred_radio_sales,
+      xlab = "Observed Sales (in '000 units)",
+      ylab = "Predicted Sales (in '000 units)")
> abline(a = 0, b = 1, col = "red")
>
```




```
> pred_news_sales <- predict(model_news, newdata = AdData_p)
> plot(AdData_p$Sales, pred_news_sales,
+      xlab = "Observed Sales (in '000 units)",
+      ylab = "Predicted Sales (in '000 units)")
> abline(a = 0, b = 1, col = "red")
>
```




```
> predict(model_tv, data.frame(TV = 151.070),
+         interval = "confidence", level = .95)
       fit      lwr      upr
1 15.24018 14.88794 15.59242
> predict(model_tv, data.frame(TV = 151.070),
+         interval = "prediction", level = .95)
       fit      lwr      upr
1 15.24018 10.77351 19.70685
> predict(model_tv, data.frame(TV = 221.380),
+         interval = "confidence", level = .95)
     fit      lwr      upr
1 19.166 18.70205 19.62994
> predict(model_tv, data.frame(TV = 221.380),
+         interval = "prediction", level = .95)
     fit      lwr      upr
1 19.166 14.68913 23.64286
```