Question 5.1

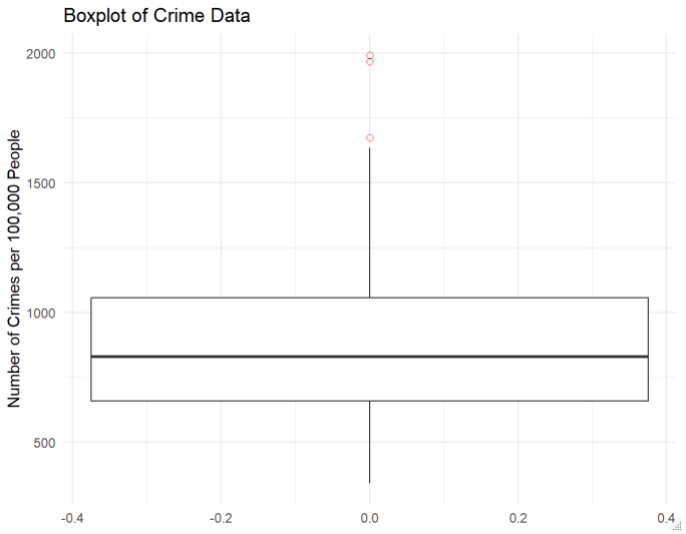
Using crime data from the file uscrime.txt (http://www.statsci.org/data/general/uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the grubbs.test function in the outliers package in R.

R code:

```
install.packages("ggplot2")

install.packages("outliers")

library(ggplot2)

library(outliers)



grubbs_result <- grubbs.test(data$Crime)

print(grubbs_result)

ggplot(data, aes(y = Crime)) +

 geom_boxplot(outlier.colour = "red", outlier.shape = 1, outlier.size = 2) +  # use red to mark outliers

 labs(title = "Boxplot of Crime Data", y = "Number of Crimes per 100,000 People") +  # graph title and lable of y

 theme_minimal()
```

Answer discussion:

```
        Grubbs test for one outlier

data:  data$Crime
G = 2.81287, U = 0.82426, p-value = 0.07887
alternative hypothesis: highest value 1993 is an outlier
```



G = 2.81287, p-value = 0.07887. According to the results of Grubbs' test, the p-value = 0.07887 is greater than the common significance level (0.05), which means we cannot reject the null hypothesis (that there are no outliers in the data). Therefore, at this significance level, there are no significant outliers in the last column of the crime data.

However, based on our dataset, the "Crime" column is univariate, which means we can use a boxplot to show the distribution of the data and mark the values that fall outside the normal range. The data point in **1993** (as highlighted by Grubbs' Test) is clearly marked in the boxplot, indicating that this data point does indeed appear unusual compared to other data points.

Question 8.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.

Answers:

Situation: We are the management team for Publix in the Atlanta midtown, and we would like to understand the factors that influence total daily grocery sales.

Potential Predictors.

Variable 1: Weekday (Monday through Sunday)

Weekday and weekend sales will vary due to the different schedules and travel habits of customers in the neighborhood.

Variable 2: Number of customers

Supermarket customer traffic directly affects the sales of groceries. The more customers patronize the store, the higher the sales of groceries.

Variable 3: Number of discounted items

Promotions on goods tend to increase the incentive for customers to buy goods, so checking the number of discounted goods items in a category can help us examine the impact of discounting activities on total sales.

Variable 4: Whether or not it is a school holiday (binary: holiday = 1 non-holiday = 0)

Considering that the households around publix are mainly school teachers as well as students, school holidays may affect our customer traffic and hence sales.

Question 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data:

M = 14.0
So = 0
Ed = 10.0
Po1 = 12.0
Po2 = 15.5
LF = 0.640
M.F = 94.0
Pop = 150
NW = 1.1
U1 = 0.120
U2 = 3.6
Wealth = 3200
Ineq = 20.1
Prob = 0.04
Time = 39.0

Show your model (factors used and their coefficients), the software output, and the quality of fit.

Note that because there are only 47 data points and 15 predictors, you'll probably notice some overfitting. We'll see ways of dealing with this sort of problem later in the course.

Answer:

In this question, there are 47 data points and 15 predictors, the description of these predictors are shown as table 1:

Table 1 The description of variables

Variable	Description
M	percentage of males aged 14–24 in total state population
So	indicator variable for a southern state
Ed	mean years of schooling of the population aged 25 years or over
Po1	per capita expenditure on police protection in 1960
Po2	per capita expenditure on police protection in 1959
LF	labour force participation rate of civilian urban males in the age-group 14-24
M.F	number of males per 100 females
Pop	state population in 1960 in hundred thousand
NW	percentage of nonwhites in the population

U1	unemployment rate of urban males 14–24
U2	unemployment rate of urban males 35–39
Wealth	wealth: median value of transferable assets or family income
Ineq	income inequality: percentage of families earning below half the median income
Prob	probability of imprisonment: ratio of number of commitments to number of offenses
Time	average time in months served by offenders in state prisons before their first release
Crime	crime rate: number of offenses per 100,000 population in 1960

To predict the crime rate with a new dataset of the 15 variables in the city, we need to find the correlation between the crime rate and 15 variables. We will use lm function and R to solve the question.

R code：

```
url <- "http://www.statsci.org/data/general/uscrime.txt"

crime_data <- read.table(url, header = TRUE)

crime_model <- lm(Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + Wealth + Ineq + Prob + Time, data = crime_data)

summary(crime_model)
```

First, we load the data and fit the data in a linear regression model.
We set the “crime” as dependent variable and “M, So, Ed, Po1, Po2, LF, M.F, Pop, NW, U1, U2, Wealth, Ineq, Prob, Time” as independent variables. Use linear model(lm) to find the correlation of the model. **Here is the output of software.**
The residuals are shown as table 2,

Table 2 The residuals

Min	1Q	Median	3Q	Max
-395.7	-98.09	-6.69	112.99	512.67

The indicators are shown as table 3,

Table 3 The indicators

Multiple R-squared	0.80
Adjusted R-squared	0.71
F-statistic	8.429
P-value	0.0000003539

The indicators show how well the model fit the data.
The R-squared is 0.8, which means the model explain 80% of the variation in the crime rate. It is high, suggesting the model fits the data fairy well. The adjusted R-squared is considering the potential overfitting, but the adjusted is still high.
The F-statistic is 8.429, which means statistically significant.
The P-value is 0.0000003539, which means some of the predictors are significantly related to the crime rate.

The factors and coefficients are shown as table 4,

Table 4 The factors and coefficients of multiple linear regression model

	Estimate	Std. Error	t value	**Pr(>|t|)**	
(Intercept)	**-5984.00**	1628.00	-3.68	**0.00**	***
M	**87.83**	41.71	2.11	**0.04**	*
So	**-3.80**	148.80	-0.03	**0.98**	
Ed	**188.30**	62.09	3.03	**0.00**	**
Po1	**192.80**	106.10	1.82	**0.08**	.
Po2	**-109.40**	117.50	-0.93	**0.36**	
LF	**-663.80**	1470.00	-0.45	**0.65**	
M.F	**17.41**	20.35	0.86	**0.40**	
Pop	**-0.73**	1.29	-0.57	**0.57**	
NW	**4.20**	6.48	0.65	**0.52**	
U1	**-5827.00**	4210.00	-1.38	**0.18**	
U2	**167.80**	82.34	2.04	**0.05**	.
Wealth	**0.10**	0.10	0.93	**0.36**	
Ineq	**70.67**	22.72	3.11	**0.00**	**
Prob	**-4855.00**	2272.00	-2.14	**0.04**	*
Time	**-3.48**	7.17	-0.49	**0.63**	

The “Estimate” column means the value of the coefficient for each variable, it shows the effect of that variable on the crime rate. The “Pr(>|t|)” column means the p-value, it shows whether the predictor is statistically significant. (Predictors with p-value less than 0.05 are generally considered statistically significant. So, the **“M, Ed, Ineq, Prob”** are statistically significant. **“U2”** is marginally significant. They show strong positive relationships with crime rates. **“Prob”** shows a strong negative effect, meaning a higher chance of getting caught reduces crime rates. Other variables like **“So, Pop, LF”** could potentially be removed in the modeling. **Other** show weak relationships with crime rates.

We **visualize** the actual crime vs the predicted crime rate with “ggplot2” function.

The figure 1 shows the result.

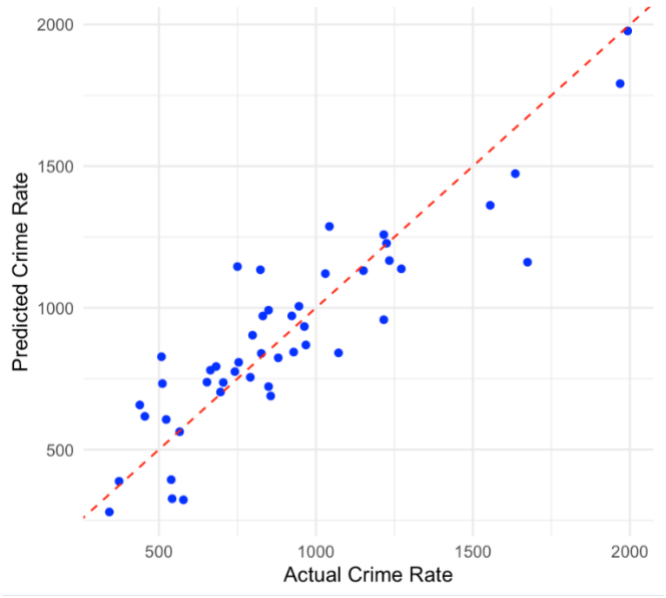


Figure 1 The actual VS the predicted

R code:

```
actual_values <- crime_data$Crime

ggplot(crime_data, aes(x = actual_values, y = fitted_values)) +

 geom_point(color = "blue") +

 geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +

 ggtitle("Actual vs Predicted Crime Rates") +

 xlab("Actual Crime Rate") +

 ylab("Predicted Crime Rate") +

 theme_minimal()
```

Next, based on the linear model and correlation, we predict the crime rate of the new dataset.

R code:

```
new_city <- data.frame(M = 14.0, So = 0, Ed = 10.0, Po1 = 12.0, Po2 = 15.5, LF = 0.640,

                       M.F = 94.0, Pop = 150, NW = 1.1, U1 = 0.120, U2 = 3.6,

                       Wealth = 3200, Ineq = 20.1, Prob = 0.04, Time = 39.0)
```

predicted_crime <- predict(crime_model, new_city)

predicted_crime

The predicted crime rate is 155.4349.