Note #04: Multiple Linear Regression (Review)

Cheng Peng

Contents

1 Introduction **1**

2 The Practical Question **1**

3 The Process of Building A Multiple Linear Regression Model **2**
3.1 Assumptions of MLR 2
3.2 The Structure of MLR 2
3.3 More on Model Specifications 4
3.4 Estimation of Regression Coefficients 5
3.5 Model Diagnostics 5
3.6 Goodness-of-fit and Variable Selection 6

4 Case Study 1 **7**

5 Case Study 2 **11**

1 Introduction

We discussed the relationship between variables in the previous two modules. The continuous variable with a normal distribution is called the response (dependent) variable and the other variable is called the explanatory (predictor, independent, or risk) variable. If the predictor variable is a factor variable, the model is called the ANOVA model which focuses on comparing the means across all factor levels. If the predictor variable is **continuous**, the model is called simple linear regression (SLR). Note that all predictor variables are assumed to be non-random.

2 The Practical Question

Maximum mouth opening (MMO) is also an important diagnostic reference for dental clinicians as a preliminary evaluation. Establishing a normal range for MMO could allow dental clinicians to objectively evaluate the treatment effects and set therapeutic goals for patients performing mandibular functional exercises.

To study the relationship between maximum mouth opening and measurements of the lower jaw (mandible). A researcher randomly selected a sample of 35 subjects and measured the dependent variable, maximum mouth opening (MMO, measured in mm), as well as predictor variables, mandibular length (ML, measured in mm), and angle of rotation of the mandible (RA, measured in degrees) of each of the 35 subjects.

The question is the maximum mouth opening (MMO) is determined by **two variables simultaneously**. We want to assess how these two variables (ML and RA) impact MMO **simultaneously**.

If we pick one predictor variable at a time, ML, to build a simple linear regression model and ignore the other predictor variable (RA), you only get the marginal relationship between MMO and ML since you implicitly assume that the relationship between MMO and ML will not be impacted by RA. This implicit assumption

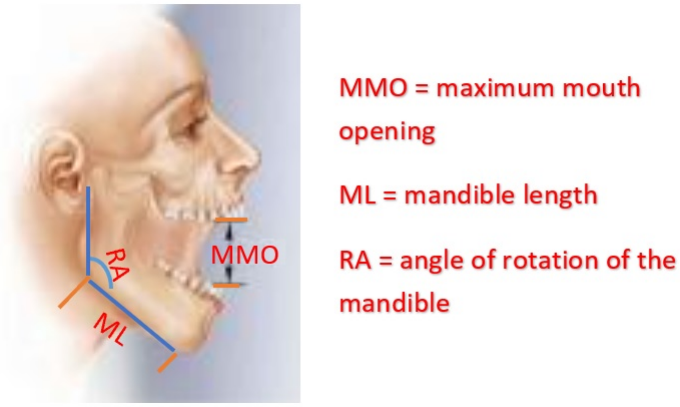


Figure 1: MMO, ML, and RA

is, in general, incorrect. We need to consider all predictor variables at the same time. This is the motivation for studying multiple linear regression (MLR).

3 The Process of Building A Multiple Linear Regression Model

The previous motivation example involves two continuous predictor variables. In real-world applications, it is common to have many predictor variables. Predictor variables are also assumed to be non-random. They could be categorical, continuous, or discrete. In a specific application, you may have a set of categorical, continuous, and discrete predictor variables in one data set.

3.1 Assumptions of MLR

There are several assumptions of multiple linear regression models.

- The response variable is a normal random variable and its mean is influenced by explanatory variables but not the variance.
- The explanatory variables are assumed to be non-random.
- The explanatory variables are assumed to be uncorrelated to each other.
- The functional form of the explanatory variables in the regression model is correctly specified.
- The data is a random sample taken independently from the study population with a specified distribution.

Some of these assumptions will be used directly to define model diagnostic measures. The idea is to assume all conditions are met (at least temporarily) and then fit the model to the data set.

3.2 The Structure of MLR

Assume that there are p predictor variables $\{x_1, x_2, \cdots, x_p\}$, the first-order linear regression is defined in the following form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

β_0 is the intercept, $\beta_1, \beta_2, \cdots, \beta_p$ are called slope parameters. if $\beta_i = 0$, the associated predictor variable x_i is uncorrelated with response vararible y. If $\beta_i > 0$, then y and x_i are positively correlated. In fact, β_1 is the increment of y as x_i increases one unit and other predictors remain unchanged.

MMO (Y)	ML (X_1)	RA (X_2)	MMO (Y)	ML (X_1)	RA (X_2)
52.34	100.85	32.08	50.82	90.65	38.33
51.90	93.08	39.21	40.48	92.99	25.93
52.80	98.43	33.74	59.68	108.97	36.78
50.29	102.95	34.19	54.35	91.85	42.02
57.79	108.24	35.13	47.00	104.30	27.20
49.41	98.34	30.92	47.23	93.16	31.37
53.28	95.57	37.71	41.19	94.18	27.87
59.71	98.85	44.71	42.76	89.56	28.69
53.32	98.32	33.17	51.88	105.85	31.04
48.53	92.70	31.74	42.77	89.29	32.78
51.59	88.89	37.07	52.34	92.58	37.82
58.52	104.06	38.71	50.45	98.64	33.36
62.93	98.18	43.89	43.18	83.70	31.93
57.62	91.01	41.06	41.99	88.46	28.32
65.64	96.98	41.92	39.45	94.93	24.82
52.85	97.85	35.25	38.91	96.81	23.88
64.43	96.89	45.11	49.10	93.13	36.17
57.25	98.35	39.44			

Figure 2: Dental Data for the multiple linear regression model (MLR)

The response variable is assumed to be a normal random variable with constant variance. If the first-order linear regression function is correct, then

$$y \to N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \sigma^2).$$

This also implies that $\epsilon \to N(0, 1)$. The residual of each data point can be estimated from the data with an assumed linear regression model.

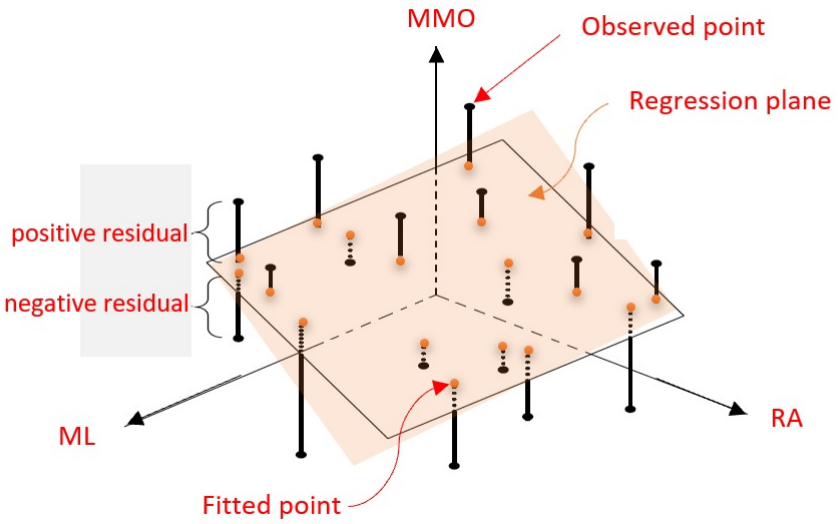


Figure 3: Illustrative regression plane: MMO vs ML and RA

For ease of illustration, let's consider the case of the MLR with two predictor variables in the motivation

example.

$$MMO = \beta_0 + \beta_1 ML + \beta_2 RA + \epsilon$$

is the first-order linear regression model. The following figure gives the graphical annotations of the fundamental concepts in linear regression models. This is a generalization of the regression line (see the analogous figure in the previous module for the simple linear regression model).

Since MMO is a normal random variable with constant variance, $MMO \to N(\beta_0 + \beta_1 ML + \beta_2 RA, \sigma^2)$, or equivalently, $\epsilon \to N(0, \sigma^2)$. The residuals are defined to be the directional vertical distances between the observed points and the regression plane.

In some practical applications, we may need **the second-order** linear regression model to reflect the actual relationship between predictor variables and the response variable. For example,

$$MMO = \alpha_0 + \alpha_1 ML + \alpha_2 RA + \alpha_3 ML^2 + \alpha_4 RA^2 + \alpha_5 ML \times RA + \epsilon$$

is called (the second-order) linear regression model. With the second-order terms in the regression function, we obtain the regression surface as shown in Figure.

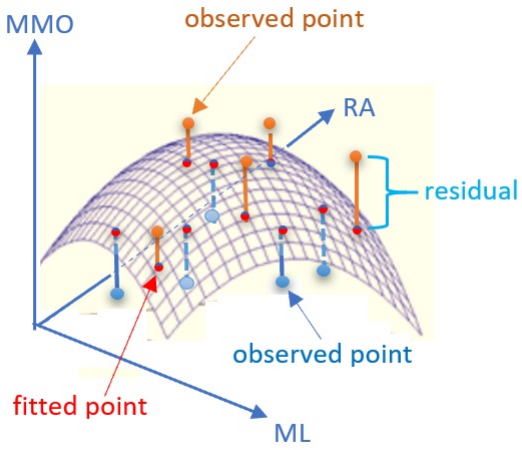


Figure 4: Illustrative regression surface: MMO vs ML and RA

If the second-order linear regression is appropriate, then $\epsilon \to N(0, \sigma^2)$ and $E[MMO] = \alpha_0 + \alpha_1 ML + \alpha_2 RA + \alpha_3 ML^2 + \alpha_4 RA^2 + \alpha_5 ML \times RA$. The residuals of the second-order linear regression model are defined to be the directional distance between the observed points and the regression surface.

3.3 More on Model Specifications

In the above section, we introduced both first- and second-order polynomial regression models. In general, it is not common to use high-order polynomial regression models in real-world applications.

- **Interaction effect** - it is common to include interaction terms (i.e., the cross product of two or more predictor variables) in the multiple linear regression models when the effect of one variable on the response variable is dependent on the other predictor variable. In other words, the interaction terms capture the **joint effect** of predictor variables. **It is rare to have third-order or higher-order interaction terms in a regression model**.
- **Dummy variables** - All categorical predictor variables are automatically converted into dummy variables (binary indicator variables). If categorical variables in the data are numerically coded, we have to turn these numerically coded variables into factor variables in the regression model.

- **Discretization and Regrouping** - Discretizing numerical predictor variables and regrouping categorical or discrete predictor variables are two basic pre-process procedures that are actually very common in many practical applications.
  - Sometimes these two procedures are required to satisfy certain model assumptions. For example, if a categorical variable has a few categories that have less than 5 observations, the resulting p-values based on certain hypothesis tests will be invalid. In this case, We have to regroup some of the categories in **meaningful ways** to resolve the **sparsity** issues in order to obtain valid results.
  - In many other applications, we want the model to be easy to interpret. Discretizing numerical variables is common. For example, we can see grouped ages and salary ranges in different applications.

3.4 Estimation of Regression Coefficients

A simple and straightforward method for estimating the coefficients of linear regression models is to minimize the sum of the squared residuals - least square estimation (LSE). To find the LSE of the regression coefficients, we need to

- choose the (first-order, second-order, or even high-order) regression function (see 3D hyper-plane or hyper-surface in the above two figures as examples).
- find the distances between the observed points and the hyper-plane (or hyper-surface). These distances are the residuals of the regression - which is dependent on the regression coefficients.
- calculate the sum of squared residuals. This sum of the residuals is still dependent on the regression coefficients.
- find the values for the regression coefficients that minimize the sum of the squared residuals. These values are called the least square estimates (LSEs) of the corresponding regression coefficients.

R function **lm()** implements the above LSE algorithm to find the regression coefficients. We have used this function in ANOVA and simple linear regression models.

3.5 Model Diagnostics

Unlike simple linear regression models, the primary assumptions of the regression model focus on the normal distribution of the response variable and the correct regression function. For multiple linear regression models, we need to impose a couple of assumptions in addition to those in the simple linear regression models

- **Residual Diagnostics**

One of the fundamental assumptions of linear regression modeling is that the response variable is normally distributed with a constant variance. This implies $\epsilon \to N(0, \sigma^2)$.

After obtaining the LSE of the regression coefficients, we can estimate the residuals and use these estimated residuals to detect the potential violations of the normality assumption of the response variable. To be more specific, we consider the first-order polynomial regression, the estimated residual of i-th observation is defined to be $e_i = MMO - \hat{\beta}_0 + \hat{\beta}_1 ML + \hat{\beta}_2 RA$

If there is no violation of the normality assumption, we would expect the following residual plot and Q-Q plot.

Some of the commonly seen poor residual plots represent different violations of various assumptions. We can try to use various transformations (such Box-Cox power transformations) of the response variable to correct the issue.

- **Multicollinearity**

Some of the predictor variables are linearly correlated. The consequence of multi-collinearity causes to unstable LSE of the regression coefficients (i.e., the LSEs of the regression coefficients are sensitive to a small

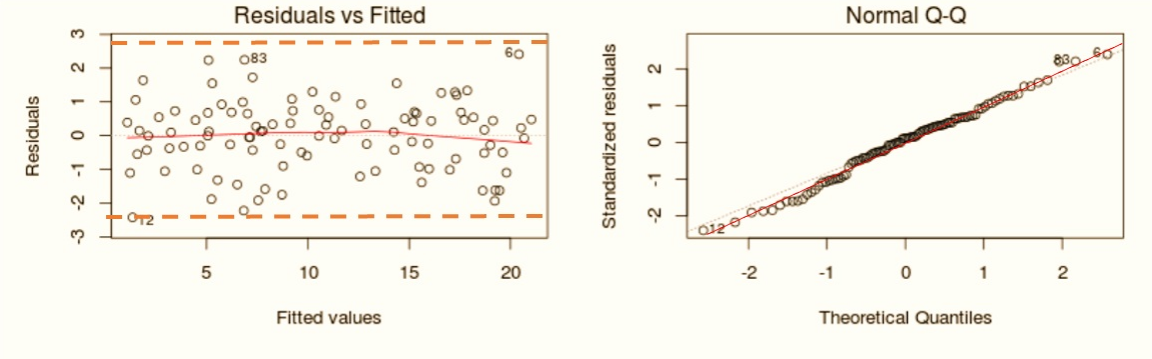


Figure 5: Good residual plot and normal Q-Q plot

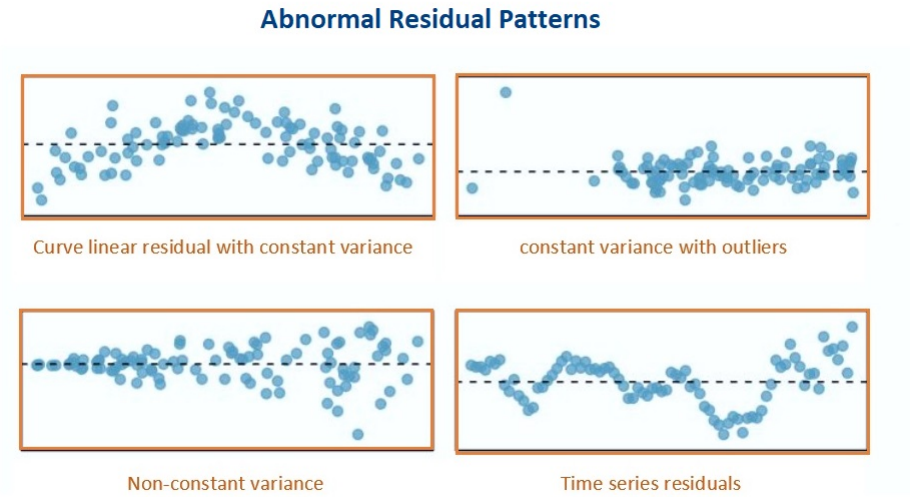


Figure 6: Poor residual plots representing various violations of the model assumptions

change in the model). It also reduces the precision of the estimate coefficients and, hence, the p-values are not reliable.

Multicollinearity affects the coefficients and p-values, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics. If our primary goal is to make predictions, we don't need to understand the role of each independent variable and we don't need to reduce severe multicollinearity.

If the primary goal is to perform association analysis, we need to reduce collinearity since both LSE and p-values are the keys to association analysis.

To detect multicollinearity, we can use the variance inflation factor (VIF) to inspect the multicollinearity of the individual predictor variable. There are some different methods to reduce multicollinearity. Centering predictor variables is one of them and works well sometimes. Some other advanced modeling-based methods are covered in more advanced courses.

3.6 Goodness-of-fit and Variable Selection

There several different goodness-of-fit measures are available for the linear regression model due to the assumption of the normality assumption of the response variable.

- **Coefficient of Determination**

We only introduce **the coefficient of determination** R^2 which measures the percentage of variability within the -values that can be explained by the regression model. In simple linear regression models, **the coefficient of determination** R^2 is simply the square of the sample Pearson correlation coefficient.

- **Statistical Significance and Practical Importance**

A small p-value of the significant test for a predictor variable indicates the variable is statistically significant but may not be practically important. On the other hand, some practically important predictor variables may not achieve statistical significance due to the limited sample size. In the practical applications, **we may want to include some of the practically important predictor variables in the final model regardless of their statistical significance**.

- **Model Selection**

One of the criteria for assessing the goodness-of-fit is the parsimony of the model. A parsimonious model is a model that accomplishes the desired level of explanation or prediction with as few predictor variables as possible. There are generally two ways of evaluating a model: Based on predictions and based on goodness of fit on the current data such as R^2 and some likelihood-based measures.

R has an automatic variable selection procedure, **step()**, which uses the goodness-of-fit measure AIC (Akaike Information Criterion) which is not formally introduced in this class due to the level of mathematics needed in the definition, but we can still use it to perform the automatic variable selection. This tutorial gives detailed examples on how to use **step()** (link).

4 Case Study 1

We use the dental data in the motivation example for the case study.

```
MMO=c(52.34, 51.90, 52.80, 50.29, 57.79, 49.41, 53.28, 59.71, 53.32, 48.53, 51.59,
      58.52, 62.93, 57.62, 65.64, 52.85, 64.43, 57.25, 50.82, 40.48, 59.68, 54.35,
      47.00, 47.23,  41.19, 42.76, 51.88, 42.77, 52.34, 50.45, 43.18, 41.99, 39.45,
      38.91, 49.10)
##
ML=c(100.85, 93.08, 98.43, 102.95, 108.24, 98.34, 95.57, 98.85,98.32, 92.70, 88.89,
     104.06,  98.18, 91.01, 96.98, 97.85, 96.89, 98.35, 90.65, 92.99, 108.97, 91.85,
     104.30, 93.16, 94.18, 89.56, 105.85, 89.29, 92.58, 98.64, 83.70, 88.46, 94.93,
     96.81, 93.13)
##
RA = c(32.08, 39.21, 33.74, 34.19, 35.13, 30.92, 37.71, 44.71, 33.17, 31.74, 37.07,
       38.71, 43.89, 41.06, 41.92, 35.25, 45.11, 39.44, 38.33, 25.93, 36.78, 42.02,
       27.20, 31.37, 27.87, 28.69, 31.04, 32.78, 37.82, 33.36, 31.93, 28.32, 24.82,
       23.88, 36.17)

DentalData = as.data.frame(cbind(MMO = MMO, ML = ML, RA = RA))
```

- **Pair-wise Scatter Plot**

This pairwise scatter plot tells whether there are significant correlations between **numerical predictor variables**.

We can see the following patterns from the above pair-wise scatter plot.

(1). Both ML and RA are linearly correlated with the response variable MMO. This is what we expected.

(2). ML and RA are not linearly correlated. This indicates that there is no collinearity issue.

(3). We also don't see any special patterns such as outliers and extremely skewed distribution. There is no need to perform discretization and regrouping procedures on the predictor variables.

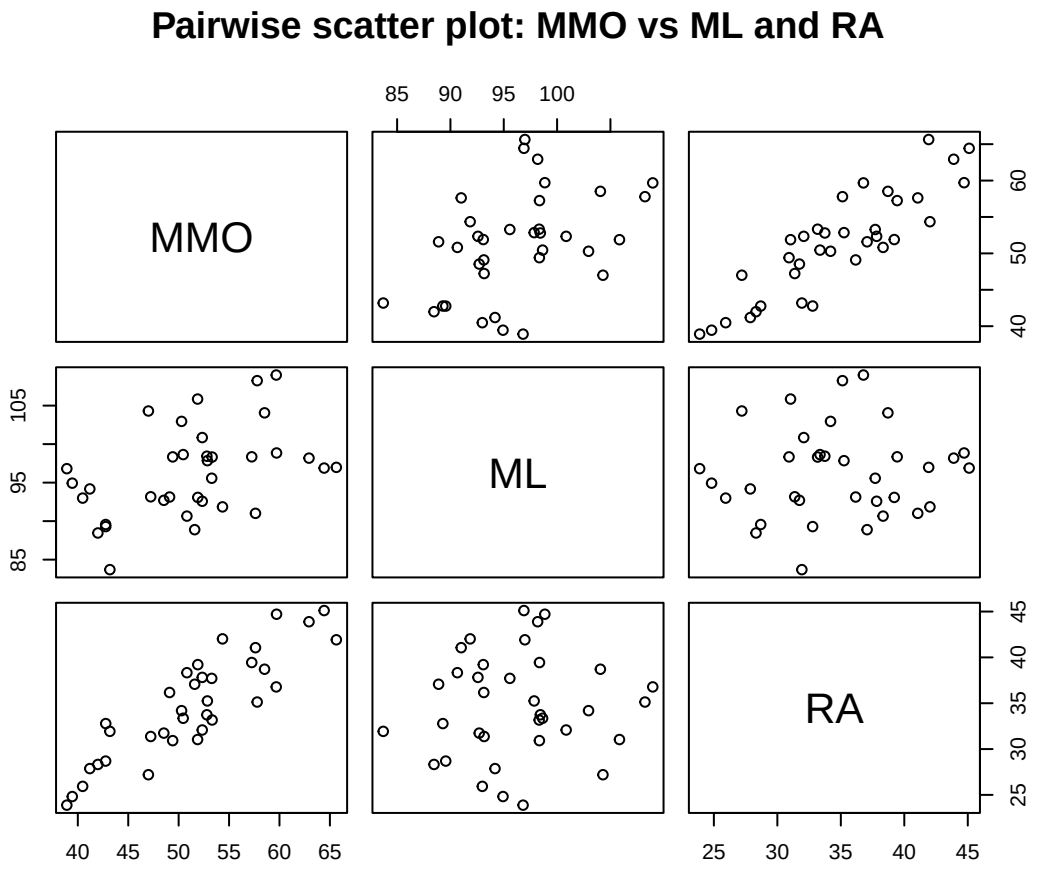


Figure 7: Pair-wise scatter plot

(4). In this data set, there is no categorical variables or categorical variable with a numerical coding system in this data set. There is no need to create dummy variables.

- **Initial model**

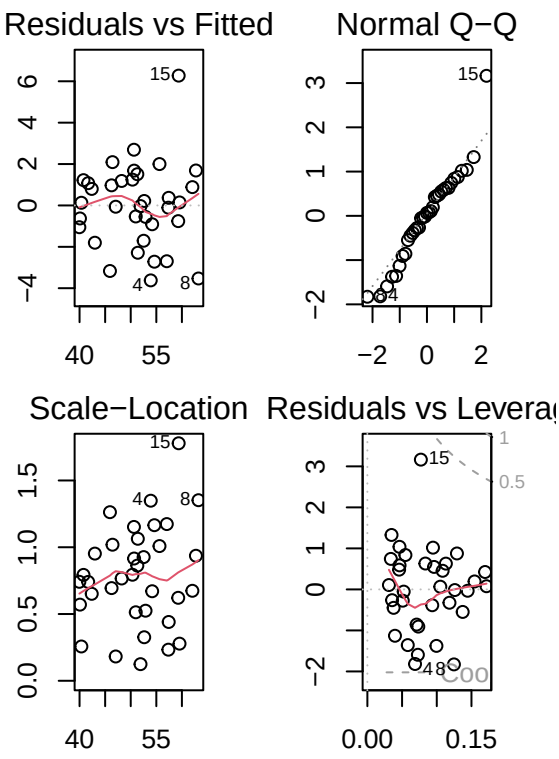
The following initial model includes all predictor variables. The residual plots indicate that

(1). One of the observations seems to be an outlier (observation 15);

(2). There is a minor violation of the assumption of constant variance.

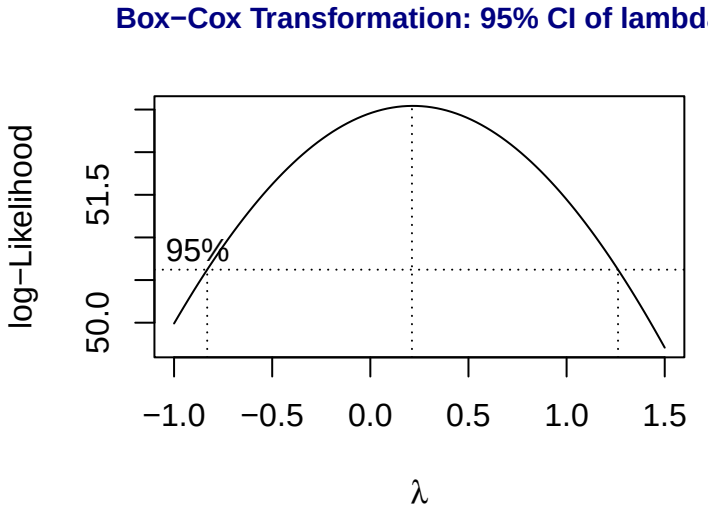
(3). There is also a minor violation of the assumption of normality of the distribution of the residuals.

```
ini.model = lm(MMO ~ ML + RA, data = DentalData)    # fit a linear model with interaction effect
par(mfrow=c(2,2), mar=c(2,3,2,2))
plot(ini.model)
```



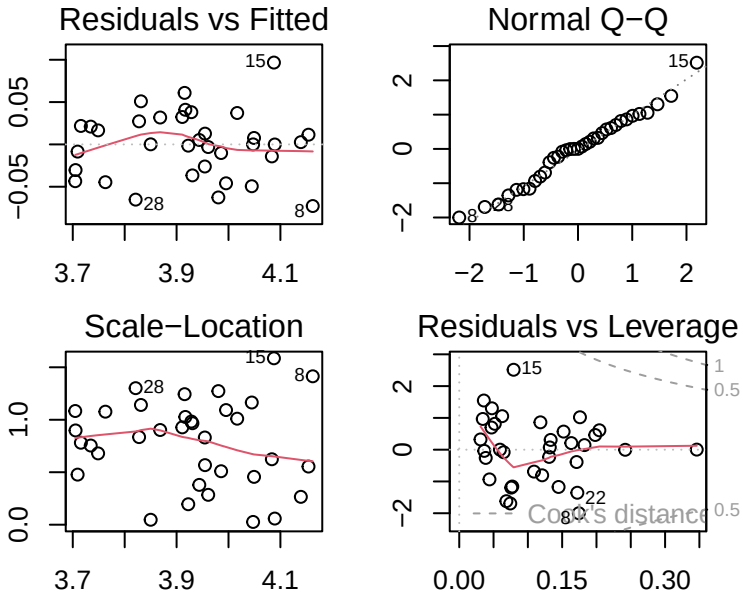
Next, we will carry the Box-Cox transformation to identify a potential power transformation of the response variable MMO.

```
library(MASS)
boxcox(MMO ~ ML + RA,
       data = DentalData,
       lambda = seq(-1, 1.5, length = 10),
       xlab=expression(paste(lambda)))
title(main = "Box-Cox Transformation: 95% CI of lambda",
      col.main = "navy", cex.main = 0.9)
```



Since both 0 and 1 are in the 95% confidence interval of λ, technically speaking, there is no need to perform the power transformation. By the optimal λ is closer to 0, we try to perform the log transformation (corresponding to $\lambda = 0$) to see whether there will some improvement of the initial model

```
transform.model = lm(log(MMO) ~ ML * RA, data  = DentalData)
par(mfrow=c(2,2), mar = c(2,2,2,2))
plot(transform.model)
```



The above residual plots indicate an improvement in model fit. We will use the transformed response to build the final model.

- **Final Model**

The model based on the log-transformed response is summarized in the following.

```
kable(summary(transform.model)$coef, caption = "Summarized statistics of the regression
      coefficients of the model with a log-transformed response")
```

Table 1: Summarized statistics of the regression coefficients of the model with a log-transformed response

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9957108	1.0023242	1.9910831	0.0553479
ML	0.0124400	0.0104503	1.1903999	0.2429256
RA	0.0296960	0.0293716	1.0110477	0.3198204
ML:RA	-0.0000884	0.0003059	-0.2890324	0.7744807

we can see that the interaction effect is insignificant in the model. We drop the highest term in the regression model either manually or automatically. In the next code chunk, we use the automatic variable selection method to find the final model.

```
transform.model = lm(log(MMO)~ML*RA, data = DentalData)
## I will use the automatic variable selection function to search the final model
final.model =  step(transform.model, direction = "backward", trace = 0)
kable(summary(final.model)$coef, caption = "Summary statistics of the regression
      coefficients of the final model")
```

Table 2: Summary statistics of the regression coefficients of the final model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.2833535	0.1176375	19.410076	0
ML	0.0094391	0.0011705	8.064482	0
RA	0.0212140	0.0012017	17.653748	0

Now we have three candidate models to select from. We extract the coefficient of determination (R^2) of each of the three candidate models.

```
r.ini.model = summary(ini.model)$r.squared
r.transfd.model = summary(transform.model)$r.squared
r.final.model = summary(final.model)$r.squared
##
Rsquare = cbind(ini.model = r.ini.model, transfd.model = r.transfd.model,
                final.model = r.final.model)
kable(Rsquare, caption="Coefficients of correlation of the three candidate models")
```

Table 3: Coefficients of correlation of the three candidate models

ini.model	transfd.model	final.model
0.9204481	0.9257218	0.9255216

The second and the third model have almost the same R^2, 92.56% and 92.57%. Both models are based on the log-transformed MMO. The interpretations of these two models are not straightforward. The initial model has a slightly lower 92.0%. Since the initial model has a simple structure and is easy to interpret, we choose the initial model as the final model to report. The summarized statistic is given in the following table.

```
summary.ini.model = summary(ini.model)$coef
kable(summary.ini.model, caption = "Summary of the final working model")
```

Table 4: Summary of the final working model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-31.4247984	6.1474668	-5.111829	1.44e-05
ML	0.4731743	0.0611653	7.735992	0.00e+00
RA	1.0711725	0.0627967	17.057792	0.00e+00

In summary, both ML and RA are statistically significant (p-value ≈ 0) and both are positively correlated to MMO. Further, for a given angle of rotation of the mandible (RA), when mandibular length (ML) increases by 1mm, the maximum mouth opening (MMO) increases by 0.473 mm. However, for holding ML, a 1-degree increase in RA will result in a 1.071 mm increase in MMO.

5 Case Study 2

We discussed the ANOVA model in module 8. In fact, the ANOVA model is a special linear regression model. The location is a factor variable. We now build a linear regression using mussel shell length as the response and the location as the predictor variable in the following (code is copied from module 8).

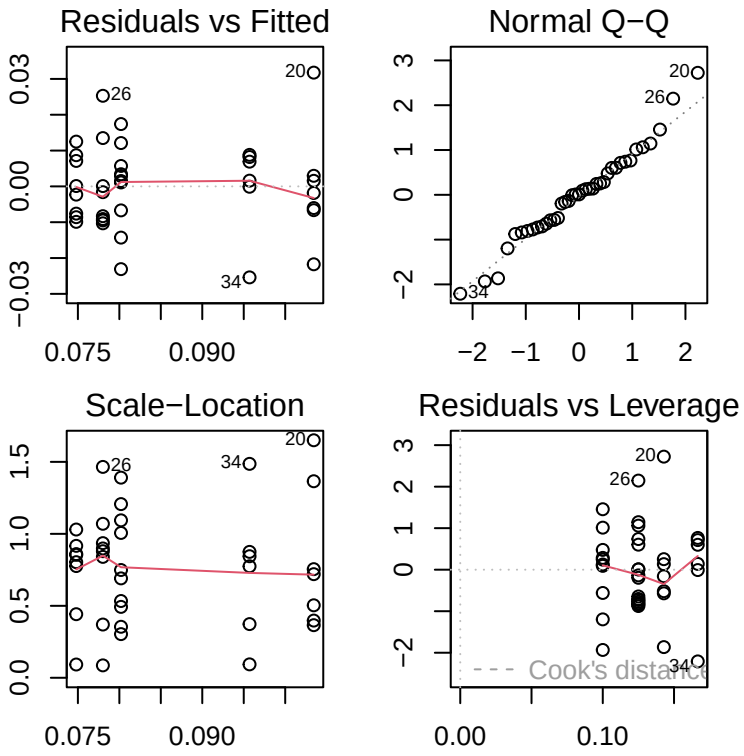
Since predictor variable location is a categorical factor variable, R function **lm()** will automatically define four dummy variables for each category except for the baseline category is, by default, the smallest character

values (alphabetical order). In our example, the value **Magadan** is the smallest. Other categories will be compared with the baseline category through the corresponding dummy variable.

To be more specific, the four dummy variables associated with the four categories will be defined by

1. locationNewport = 1 if the location is Newport, 0 otherwise;
2. locationPetersburg = 1 if the location is Petersburg, 0 otherwise;
3. locationTillamook = 1 if the location is Tillamook, 0 otherwise;
4. locationTvarminne = 1 if the location is Tvarminne, 0 otherwise.

```
x1 = c(0.0571,0.0813, 0.0831, 0.0976, 0.0817, 0.0859, 0.0735, 0.0659, 0.0923, 0.0836)
x2 = c(0.0873,0.0662, 0.0672, 0.0819, 0.0749, 0.0649, 0.0835, 0.0725)
x3 = c(0.0974,0.1352, 0.0817, 0.1016, 0.0968, 0.1064, 0.1050)
x4 = c(0.1033,0.0915, 0.0781, 0.0685, 0.0677, 0.0697, 0.0764, 0.0689)
x5 = c(0.0703,0.1026, 0.0956, 0.0973, 0.1039, 0.1045)
len  = c(x1, x2, x3, x4, x5)       # pool all sub-samples of lengths
location = c(rep("Tillamook", length(x1)),
             rep("Newport", length(x2)),
             rep("Petersburg", length(x3)),
             rep("Magadan", length(x4)),
             rep("Tvarminne", length(x5)))     # location vector matches the lengths
data.matrix = cbind(len =  len, location = location)   # data a data table
musseldata = as.data.frame(data.matrix)         # data frame
## End of data set creation
##
## Starting building the ANOVA model
anova.model.01 = lm(len ~ location, data = musseldata)  # define a model for generating the ANOVA
##
par(mfrow=c(2,2), mar = c(2,2,2,2))
plot(anova.model.01)
```



The above residual plots indicate no serious violation of the model assumption. The model that generates the above residual plot will be used as the final working model. The inference of the regression coefficients is summarized in the following table.

```
sum.stats = summary(anova.model.01)$coef
kable(sum.stats, caption = "Summary of the ANOVA model")
```

Table 5: Summary of the ANOVA model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0780125	0.0044536	17.5168782	0.0000000
locationNewport	-0.0032125	0.0062983	-0.5100593	0.6133053
locationPetersburg	0.0254304	0.0065193	3.9007522	0.0004300
locationTillamook	0.0021875	0.0059751	0.3661039	0.7165558
locationTvarminne	0.0176875	0.0068029	2.5999834	0.0136962

From the above summary tale, we can see that P-values associated with location dummy variables locationNewport, locationTillamook are bigger than 0.05 meaning the means associated with **Newport**, **Tillamook**, and the baseline **Magadan** (not appearing in the summary table). The p-values associated with **Petersburg** and **Tvarminn** are less the 0.05 which implies that the mean length of these two locations is significantly different from that of the baseline location **Magadan**. Further, the coefficient associated with dummy variable **locationPetersburg** indicates that the mean length of the mussel shell in **Petersburg** is 0.0543 units longer than that in the baseline location **Magadan**. We can also interpret the coefficients associated with **locationTvarminne**.