Week #9 Note: Correlation and Simple Linear Regression

Cheng Peng

3/25/2021

Contents

1 Introduction

In the previous module, we discussed the relationship between a continuous variable (the length of mussel shells) and a categorical variable (location, also called factor variable). If there is no association between the two variables, the means of all populations are equal. If there is an association between the continuous variable and the factor variable, then the means of the populations are not identical.

The continuous variable is assumed to normal distribution. The continuous variable is also called the response variable (dependent variable) and the factor variable is called the predictor variable (or explanatory variable).

A natural question is how to characterize the relationship between two continuous variables.

1.1 The question and the data

Example: Amyotrophic lateral sclerosis (ALS) is characterized by a progressive decline of motor function. The degenerative process affects the respiratory system. To investigate the longitudinal impact of nocturnal noninvasive positive-pressure ventilation on patients with ALS. Prior to treatment, they measured partial pressure of arterial oxygen (Pao2) and partial pressure of arterial carbon dioxide (Paco2) in patients with the disease. The results were as follows:

PTID	Paco2	Pao2	PTID	Paco2	Pao2
1	40.0	40.0	16	41.8	41.8
2	47.0	47.0	17	33.0	33.0
3	34.0	34.0	18	43.1	43.1
4	42.0	42.0	19	52.4	52.4
5	54.0	54.0	20	37.9	37.9
6	48.0	48.0	21	34.5	34.5
7	53.6	53.6	22	40.1	40.1
8	56.9	56.9	23	33.0	33.0
9	58.0	58.0	24	59.9	59.9
10	45.0	45.0	25	62.6	62.6
11	54.5	54.5	26	54.1	54.1
12	54.0	54.0	27	45.7	45.7
13	43.0	43.0	28	40.6	40.6
14	44.3	44.3	29	56.6	56.6
15	53.9	53.9	30	59.0	59.0

Figure 1: Length of clamshells

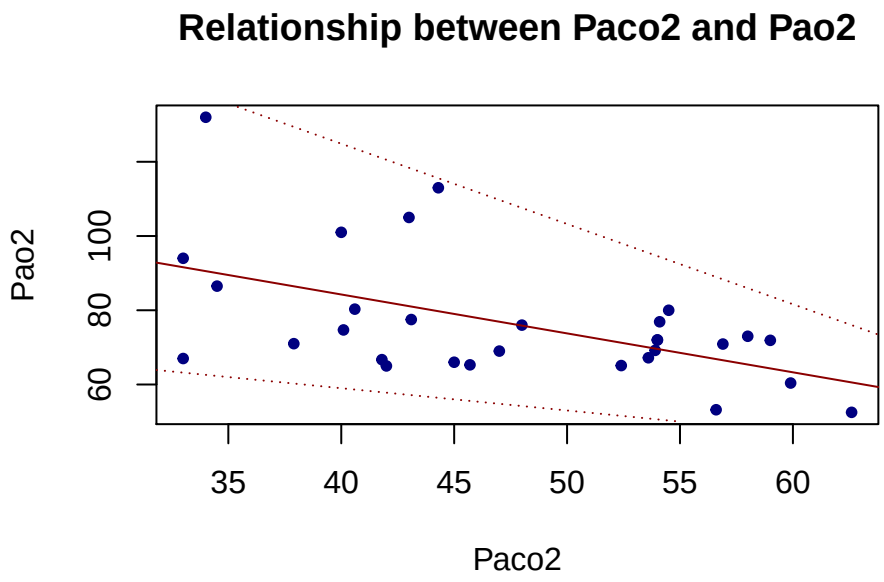
Source: M. Butz, K. H.Wollinsky, U.Widemuth-Catrinescu, A. Sperfeld, S. Winter, H. H. Mehrkens, A. C. Ludolph, and H. Schreiber, "Longitudinal Effects of Noninvasive Positive-Pressure Ventilation in Patients with Amyotrophic Lateral Sclerosis," *American Journal of Medical Rehabilitation*, 82 (2003) 597–604.

The layout of the data table is shown below. Each subject (patient ID) has one **record** with two pieces of information Paco2 and Pao2.

1.2 Visual Inspection for Association

We define two vectors to store the sample values of partial pressure of arterial oxygen (Pao2) and partial pressure of arterial carbon dioxide (Paco2). Before conducting analysis, we make a scatter plot to visualize the relationship between the two continuous variables.

```
# define the data sets based on the given data table.
Paco2 =c(40.0, 47.0, 34.0, 42.0, 54.0, 48.0, 53.6, 56.9, 58.0, 45.0, 54.5, 54.0, 43.0,
         44.3, 53.9, 41.8, 33.0, 43.1, 52.4, 37.9, 34.5, 40.1, 33.0, 59.9, 62.6, 54.1,
         45.7, 40.6, 56.6, 59.0)
Pao2 = c(101.0, 69.0, 132.0, 65.0, 72.0, 76.0, 67.2, 70.9, 73.0, 66.0, 80.0, 72.0,
         105.0, 113.0, 69.2, 66.7, 67.0, 77.5, 65.1, 71.0, 86.5, 74.7, 94.0, 60.4,
         52.5, 76.9, 65.3, 80.3, 53.2, 71.9)
## scatter plot
plot(Paco2, Pao2,
     pch = 20,
     col = "navy",
     main = "Relationship between Paco2 and Pao2",
     xlab = "Paco2",
     ylab = "Pao2"
)
## The following two lines are not required when you make this scatter plot
segments(30, 65, 55, 50, lty = 3, col = "darkred")
segments(33, 140, 70, 60, lty = 3, col = "darkred")
abline(lm(Pao2 ~ Paco2), col = "darkred")
```



We can see from the above scatter plot that there is a negative association between Paco2 and Pao2 since Pao2 decreases as Paco2 increases. We can also see that **the variance** of Pao2 is also decreasing as Paco2 increases.

How to quantify the above association?

2 The strength of linear correlation - coefficient of correlation

2.1 Definition of Pearosn correlation coefficient

One well-known quantity for measuring the **linear association** between two numerical variables is the Pearson-correlation coefficient. The sample Pearson correlation coefficient is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

2.2 Interpretation of correlation coefficient

The interpretation of the Pearson correlation coefficient is customarily given in the following

- if $r > 0$, then x and y are positively correlated; if $r < 0$, then x and y are negatively correlated;
- if $r = 0$, there is **no** linear correlation between x and y.
- if $|r| < 0.3$, there is a **weak** linear correlation between x and y.
- if $0.3 < |r| < 0.7$, there is a **moderate** linear correlation between x and y.
- if $0.7 < |r| < 1.0$, there is a **strong** linear correlation between x and y.
- if $r = 1$, there is a **perfect** linear correlation between x and y.

In R, we use command **cor()** to calculate the above Pearson correlation coefficient.

```
Pearson.correlation = cbind(r=cor(Paco2, Pao2))
kable(Pearson.correlation, caption = "Pearson correlation coefficient",
      align="c")
```

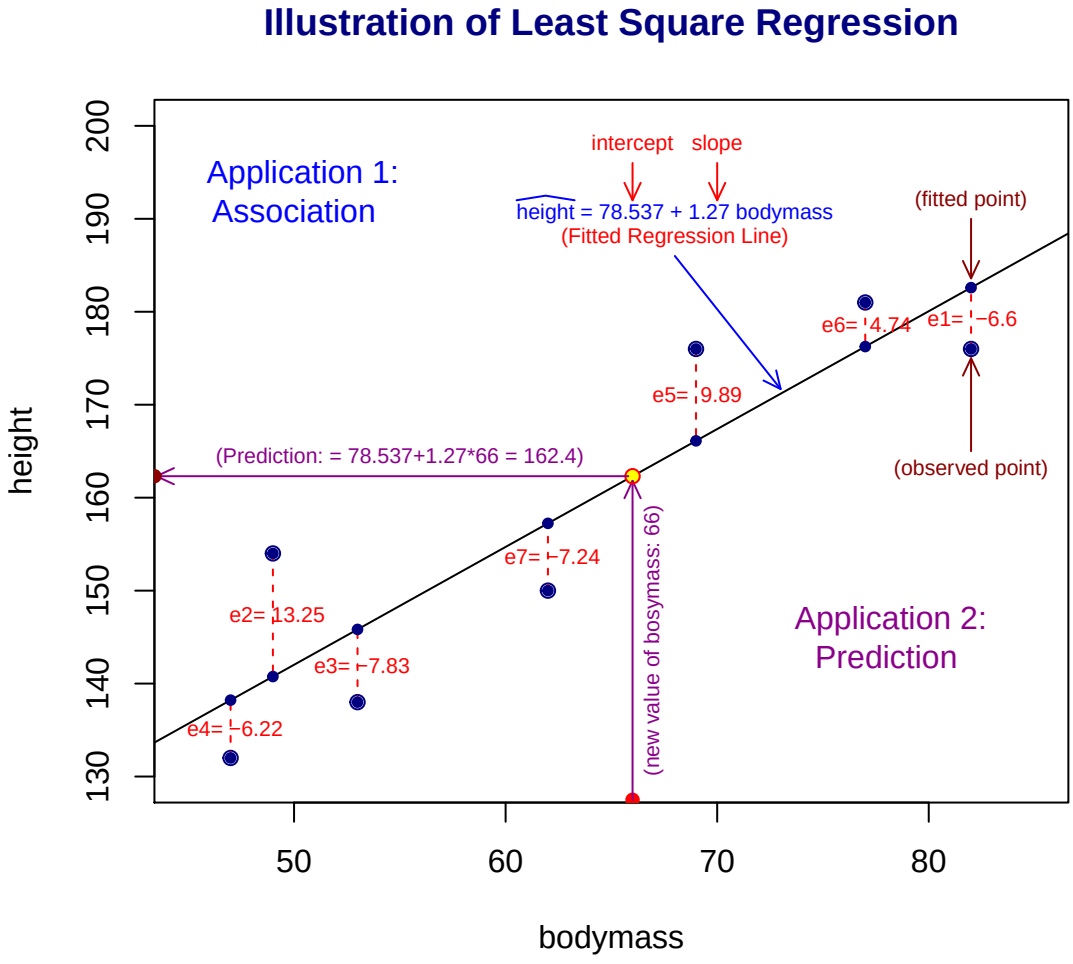
Table 1: Pearson correlation coefficient

r
-0.5307874

Therefore, there is a weak negative linear correlation between the partial pressure of arterial oxygen (Pao2) and partial pressure of arterial carbon dioxide (Paco2).

3 Least square regression: structure, diagnostics, and applications

For illustration purposes, we make the following plot based on an artificial data set to introduce several important concepts of the least square regression model.



3.1 Definitions

The following concepts are annotated on the above figure.

- The variable associated with the vertical variable is called the **response** variable. The *response** variable is always placed on the left-hand side of the model formula.
- The variable that impacts the value of the response variable is called the explanatory variable (also called predictor, independent variables).
- The points in the figure plotted based on the data set are called observed data points.
- The line in the figure is called the **fitted regression line**.
- The points on the fitted regression line are called **fitted data points**.
- The difference between coordinates observed and fitted points is called ** the residual** of the observed data point. The residual is, in fact, called the fitted error. It reflects the goodness of the fitted regression line. e_i $(i = 1, 2, \cdots, n)$ are the estimated residual errors.
- The **best** regression line is obtained by minimizing the sum of the squared errors, $e_1^2 + e_2^2 + \cdots + e_2^2$. This is the reason why we call the **best** regression line the **least square regression line** .
- The intercept and slope completely determine a straight line. The intercept and slope of the **least square** regression line are obtained by minimizing the sum of the squared residual errors.
- The statistical term *linear model* is the equation of the fitted regression line.
- There are two possible applications of a linear regression model.
  - **Association analysis** - basic describes how the change of explanatory variable impacts the value of the response variable.
  - **Prediction analysis** - predict the values of the response variable based on the corresponding **new values** of the explanatory variable.

In this module, we only discuss the least square regression with ONE explanatory variable. In the next module, we will generalize this model to multiple explanatory variables.

3.2 Assumptions of Least Square Regression

The assumptions of the least square linear regression are identical to the ones of the ANOVA.

- The response variable is a **normal random variable**. Its mean is dependent on the **non-random** explanatory variable.
- The variance of the response variable is constant (i.e., its variance is NOT dependent on the **non-random** explanatory variable).
- The relationship between the response and explanatory variables is assumed to be correctly specified.

3.3 Model Building and Diagnostics

We will use **lm()** and the artificial data used in the above plot to find the least square estimate of **parameters**: intercept and slope.

```
height <- c(176, 154, 138, 132, 176, 181, 150)
bodymass <- c(82, 49, 53, 47, 69, 77, 62)
ls.reg <- lm(height ~ bodymass)
parameter.estimates <- ls.reg$coef
kable(parameter.estimates,
      caption = "Least square estimate of the intercept and slope",
      align='c')
```

Table 2: Least square estimate of the intercept and slope

	x
(Intercept)	78.627588
bodymass	1.267897

The least square estimated intercept and slope are approximately equal to 78.63 and 1.27. Therefore, the fitted least square regression line is $\widehat{height}_i = 78.63 + 1.27 \times bodymass_i$. The residual error $e_i = height_i - \widehat{height}_i$, for $i = 1, 2, \cdots, n$.

Since the explanatory variable is implicitly assumed to be a non-random variable, we can see the relationship between the response variable and the residual in the following general representation.

$$response.variable = \alpha + \beta \times predictor.variable + \epsilon$$

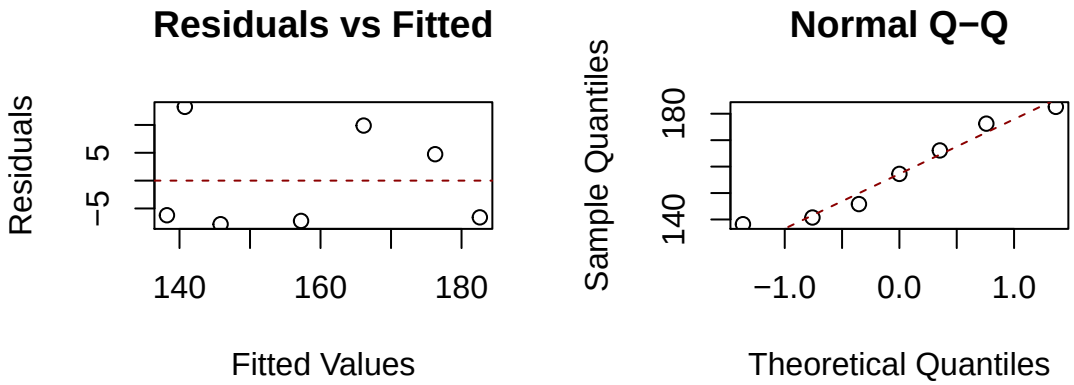
The assumption that the response variable is normal with a constant variance is equivalent to that ϵ is a normal random variable with mean 0 and constant variance σ_0^2.

The residual ϵ is estimated by the errors e_i. Therefore, we can look at the distribution of $\{e_1, e_2, \cdots, e_n\}$ to see potential violations of the model assumptions.

3.4 Residual Diagnostics and Remedies

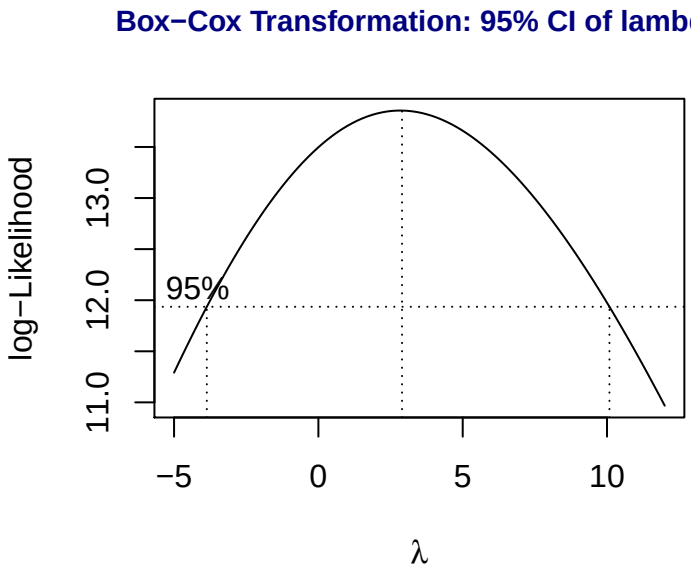
We make four default residual plots from R function **lm()** in R.

```
par(mfrow = c(1,2))  # par => graphic parameter
                     # mfrow => splits the graphic page into panels
#
plot(ls.reg$fitted.values, ls.reg$residuals,
     main="Residuals vs Fitted",       # title of the plot
     xlab = "Fitted Values",           # label of X-axis
     ylab = "Residuals"                # label of y-axis
     )
abline(h=0, lty=2, col = "darkred")
##
qqnorm(ls.reg$fitted.values, main = "Normal Q-Q")
qqline(ls.reg$fitted.values, lty = 2, col = "darkred")
```



We can see that there seems to have a minor violation of the assumption of constant variance from the residual plot in the top left figure. In statistics, we have a transformation to stabilize the variable. We will introduce the well-known Box-Cox transformation in this module. It is a generic transformation and was designed to identify the optimal power transformation to stabilize the variance and maintain the normality. Sometimes it works really well but not always. It is always worth a try in case we observed the pattern of non-constant variance.

```
library(MASS)
boxcox(height ~ bodymass, lambda = seq(-5, 12, length = 10),
       xlab=expression(paste(lambda)))
title(main = "Box-Cox Transformation: 95% CI of lambda",
      col.main = "navy", cex.main = 0.9)
```



The above plot shows the 95% confidence interval of the power (λ) on the potential power transformation of the response variable *height*.

In practice, we choose the **most convenient** number in the interval as the power to transform the response variable. Since 1 is in the interval, it is necessary to perform a power transformation. We can also choose the logarithmic transformation of height since $\lambda = 0$ is also in the interval.

Note: A special power transformation is the logarithmic transformation of the response if we choose $\lambda = 0$.

3.5 Final Model with Applications

Once the final model is identified, we can use it in two different ways: association and prediction.

In the **association analysis**, we interpret the regression coefficient associated with the significant explanatory variable.

In this toy example, the summarized statistics are extracted and summarized in the following

```
kable(summary(ls.reg)$coef, caption = "Summary of regression model")
```

Table 3: Summary of regression model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	78.627588	18.7505570	4.193347	0.0085442
bodymass	1.267897	0.2929519	4.328005	0.0075135

The p-value of testing the null hypothesis that the slope parameter to be zero is 0.0075. We reject the null hypothesis $H_0 : slope = 0$. This implies that **bodymass** impacts **height**. To be more specific, as the value of **bodymass** increases by one unit, the response variable **height** will increase by 1.27 cm.

In the **prediction analysis**, we can predict the **height** with any given new **bodymass** that is not in the data set. For example, assume **bodymass = 75**, we can then use R function **predict()** to predict the **height**.

```
pred.height = predict(ls.reg, newdata = data.frame(bodymass=75), interval = "prediction", level=0.05)
kable(pred.height, caption="The predicted height with body mass: 75")
```

Table 4: The predicted height with body mass: 75

fit	lwr	upr
173.7199	172.9821	174.4577

For a given person with a body mass of 75 unit, the predicted height of that person is 173.7cm with a 95% predictive interval [173.0, 174.5].

In summary: It is dependent on the objectives of your data analysis,

- if the objective is association analysis, the summary will focus on the interpretation of the regression coefficient with a p-value < 0.05.
- if the objective is prediction, the summary will focus on the predicted value and its predictive interval.
- if the objectives are both association analysis and prediction, then summary both results as shown above.

4 Case Study: Amyotrophic lateral sclerosis analysis revisited

We only perform the least square regression analysis about the relationship between the partial pressure of arterial oxygen (Pao2) and partial pressure of arterial carbon dioxide (Paco2) in patients with the disease.

The scatter plot of Paco2 versus Pao2 in section 1 indicates a negative association between them. Next, we assume that the linear relationship between Paco2 and Pao2 and build the least square regression in the following steps.

4.1 Objectives

The objective of this analysis is to build a linear regression model and then use this model to

- assess how Pao2 impacts Paco2.
- predict the value of Paco2 for given Pao2

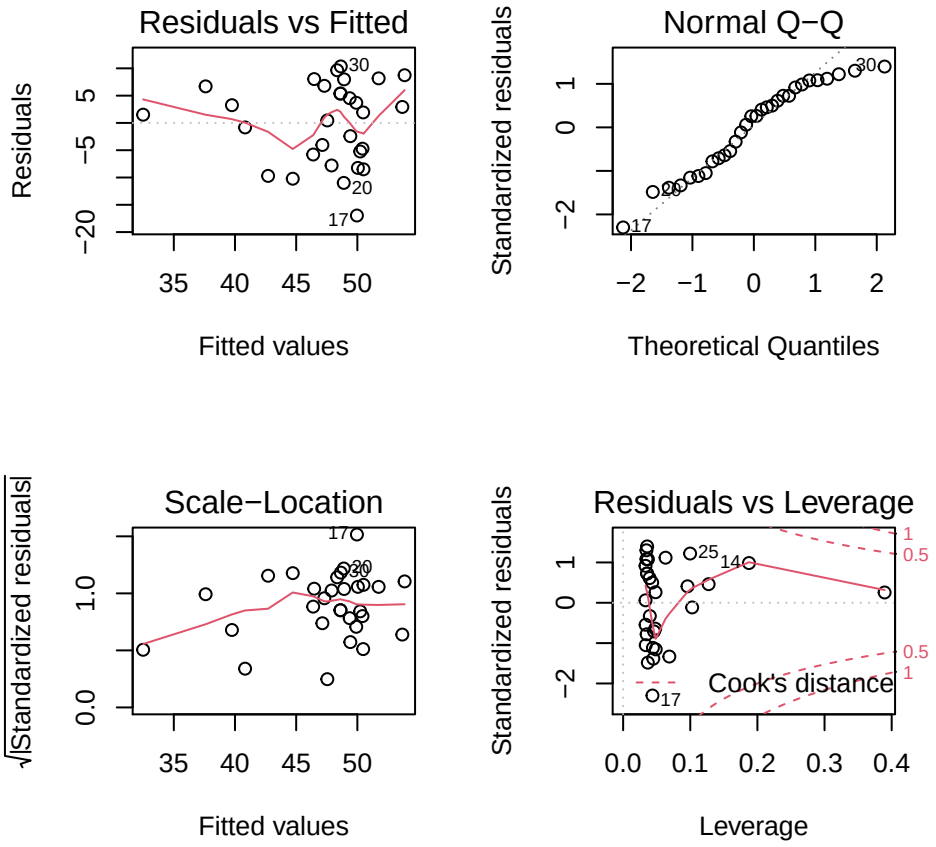
4.2 Model Fitting

We fit a least square regression to the data first. Based on the objective of the analysis, Paco2 will be the response variable and Pao2 will be the explanatory variable. We first fit the following model

$$Paco2 = \alpha + \beta \times Pao2 + \epsilon$$

α, β, and ϵ are called intercept, slope, and residuals, respectively. Then carry out the diagnostics of the above model and find the potential remedy.

```
Paco2 =c(40.0, 47.0, 34.0, 42.0, 54.0, 48.0, 53.6, 56.9, 58.0, 45.0, 54.5, 54.0, 43.0,
         44.3, 53.9, 41.8, 33.0, 43.1, 52.4, 37.9, 34.5, 40.1, 33.0, 59.9, 62.6, 54.1,
         45.7, 40.6, 56.6, 59.0)
Pao2 = c(101.0, 69.0, 132.0, 65.0, 72.0, 76.0, 67.2, 70.9, 73.0, 66.0, 80.0, 72.0,
         105.0, 113.0, 69.2, 66.7, 67.0, 77.5, 65.1, 71.0, 86.5, 74.7, 94.0, 60.4,
         52.5, 76.9, 65.3, 80.3, 53.2, 71.9)
##
ls.reg0 <- lm(Paco2 ~ Pao2)   # fitting a least square regression
par(mfrow = c(2,2))   # split the graphic page into 4 panels
plot(ls.reg0)
```



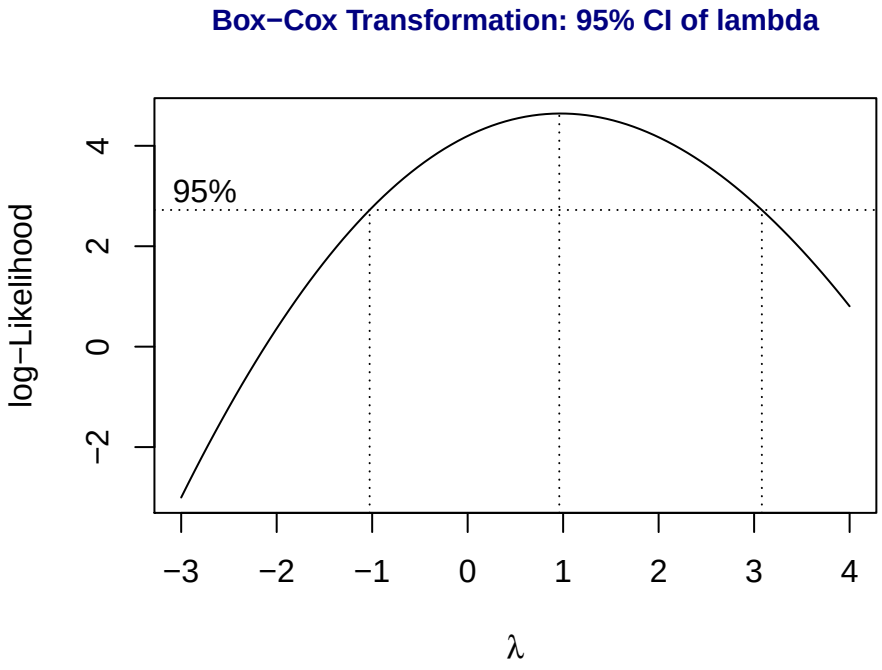
The residual diagnostic plots show that there are violations of the model assumptions. the Q-Q plot does not support the normality assumption of the residuals. The top-left residual plot does not support the constant variance assumption since the variance of the residual increases as the fitted value increases.

Next, we perform the Box-Cox transformation.

4.3 Box-Cox Transformation

In the Box-cox transformation, the range of potential λ is selected by trial-and-error so that the figure should contain the 95% confidence interval. We will use a function in the library **{MASS}**. If you don't have the library on your computer, you need to install it and then load it to the work-space.

```
library(MASS)
boxcox(Paco2 ~ Pao2, lambda = seq(-3, 4, length = 10),
       xlab=expression(paste(lambda)))
title(main = "Box-Cox Transformation: 95% CI of lambda",
      col.main = "navy", cex.main = 0.9)
```



The above Box-Cox procedure indicates that the power-transformation will not improve the residual plots. I will not try other transformations in this course and simply use the above model as the final working model for prediction and perform association analysis.

4.4 Model Applications

Recall that we will use the final working model to assess the association between the Paco2 and Pao2 and predict the value of Paco2 with new Pao2 as well.

We first present summary statistics of the least square regression model in the following table.

```
ls.reg.final <-lm(Paco2 ~ Pao2)
kable(summary(ls.reg.final)$coef,
      caption ="Summary of the final least square regression model")
```

Table 5: Summary of the final least square regression model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	67.9716010	6.3535426	10.698221	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
Pao2	-0.2687739	0.0811017	-3.314037	0.0025473

We can see that Pao2 significantly impacts Paco2 with a p-value = 0.0025. To be more specific, as Pao2 increases by a unit, the Paco2 **decreases** by about 0.27. The negative sign of the estimated slope indicates the negative linear association between Paco2 and Pao2.

Now, let assume that there two new patients who Pao2 levels 63 and 75, respectively. Note that these two Pao2 are **within the range of Pao2**.

```
## put the new observations in the form of the data frame.
new.pao2 = data.frame(Pao2 = c(63,75))
##
pred.new = predict(ls.reg.final, newdata = new.pao2,
                   interval = "prediction",
                   level = 0.05)
pred.new.cbind = cbind(Pao2.new=c(63,75), pred.new)
kable(pred.new.cbind, caption = "95% predictive intervals of Paco2")
```

Table 6: 95% predictive intervals of Paco2

Pao2.new	fit	lwr	upr
63	51.03884	50.54862	51.52906
75	47.81356	47.32818	48.29893

The above predictive table indicates that the predicted value of Paco2 with Pao2 = 63 is about 51.04 with a 95% predictive interval [50.55, 51.23]. The predicted of Paco2 is 47.81 with a 95% predictive interval [47.23, 48.30] for Pao2 = 73.