SLR

I- Introduction

(See Biometry 721-821 workbook (2024) page 43)

II- The Basics

Load required pacakges

```r
library(RcmdrMisc)
```

```
Loading required package: car
```

```
Loading required package: carData
```

```
Loading required package: sandwich
```

```r
library(ggplot2)
```

Importing data

```r
chap6data1 <- read.csv("chap6data1.csv", sep=",", header=T)
dim(chap6data1)
```

```
[1] 44  2
```

```
head(chap6data1)
```

```
  BodyWeight MetabolicRate
1       49.9          1079
2       50.8          1146
3       51.8          1115
4       52.6          1161
5       57.6          1325
6       61.4          1351
```

```
chap6data2 <- read.csv("chap6data2.csv", sep=",", header=T)
dim(chap6data2)
```

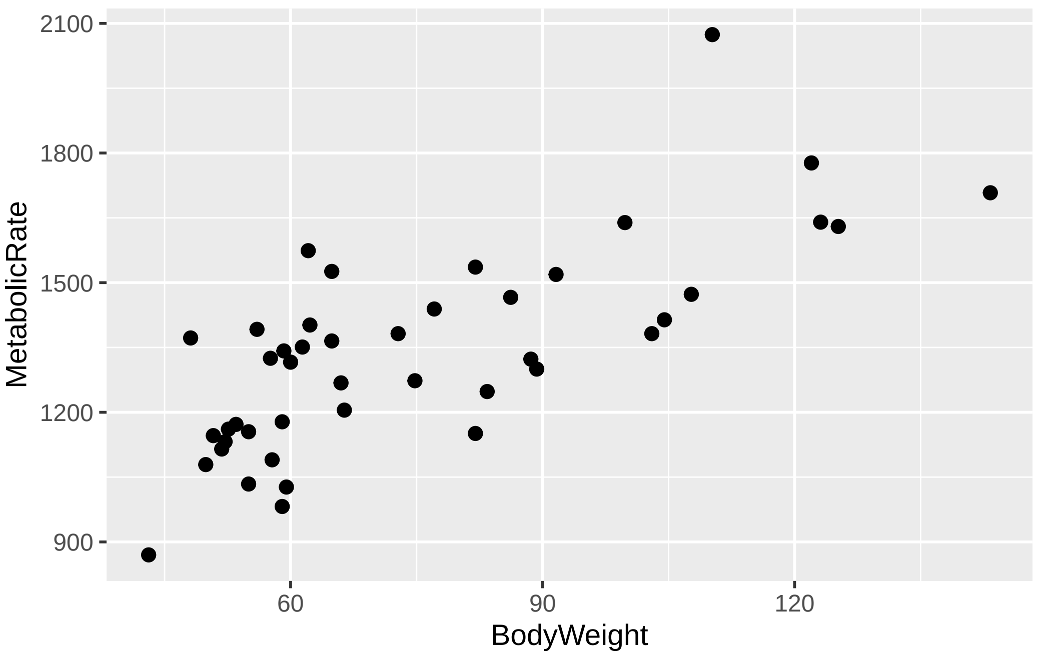
```
[1] 100   2
```

```
head(chap6data2)
```

```
  Time Mark
1   40   20
2   40   23
3   40   19
4   40   19
5   40   18
6   40   20
```

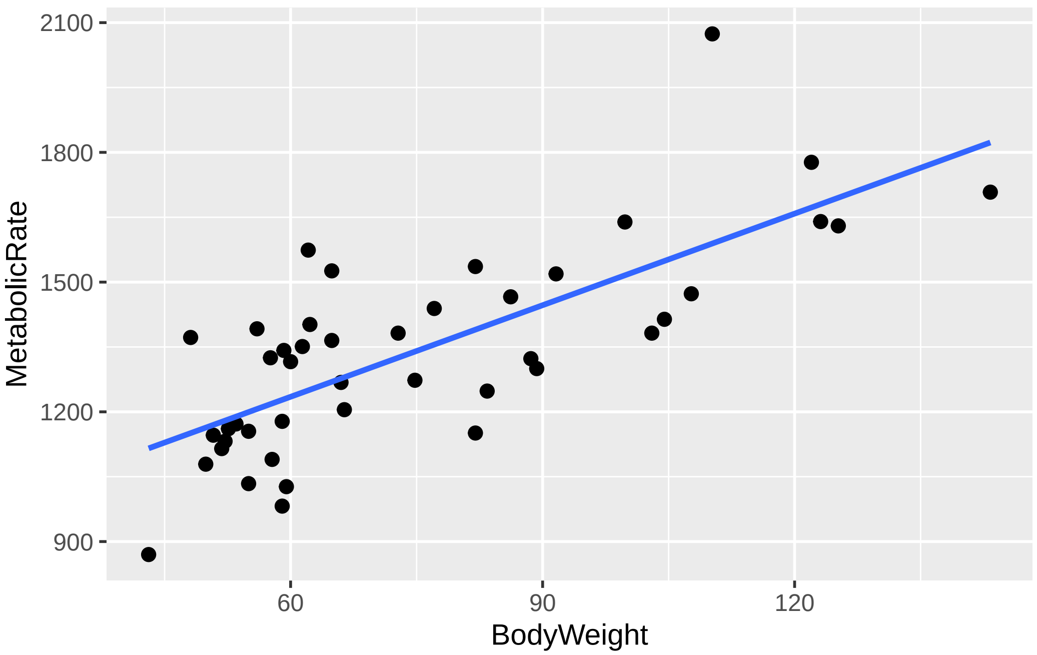
#Example 1 #visual presentation of data

```
chap6data1.plot1 <- ggplot(data=chap6data1,aes(y=MetabolicRate,x=BodyWeight))
chap6data1.plot1 + geom_point(size=2)
```



```
chap6data1.plot1 + geom_point(size=2)+ geom_smooth(method=lm,se=F)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Fit the model

```r
chap6data1.model1 <- lm(MetabolicRate~BodyWeight,data=chap6data1)
summary(chap6data1.model1)
```

```

Call:
lm(formula = MetabolicRate ~ BodyWeight, data = chap6data1)

Residuals:
    Min      1Q  Median      3Q     Max 
-245.74 -113.99  -32.05  104.96  484.81 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 811.2267    76.9755  10.539 2.29e-13 ***
BodyWeight    7.0595     0.9776   7.221 7.03e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 157.9 on 42 degrees of freedom
Multiple R-squared:  0.5539,    Adjusted R-squared:  0.5433 
F-statistic: 52.15 on 1 and 42 DF,  p-value: 7.025e-09
```

The formula of the fitted model is

$$Y = 811.2267 + 7.0595X$$

Interpretation of the parameter estimates

Intercept

The average metabolic rate when body weight is zero is approximately 811.2267 units.

BodyWeight

For each additional unit of body weight, the metabolic rate increases by about 7.0595 units.

Determine whether greater weight is associated with greater metabolism.

Given the positive coefficient (7.0595) for BodyWeight, we can conclude that greater weight is indeed associated with greater metabolism. This positive relationship suggests that as body weight increases, metabolic rate also increases.

III- Diagnostic checks

III-1- Determine and interpret the coefficient of determination of the metabolism data

Since Multiple R-squared(R^2) is equal to 0.5539, then approximately 55.39% of the variability in Metabolic Rate can be explained by Body weight.

Model diagnostics

To assess the influence of individual data points on the linear regression model `chap6data1.model1`, we use influence measures and studentized residuals.

```
influence.measures(chap6data1.model1)$infmat[1:6, ]
```

```
       dfb.1_    dfb.BdyW       dffit    cov.r       cook.d        hat
1 -0.10780127  0.08607243 -0.12020219 1.085030 0.0073475621 0.04664371
2 -0.02944633  0.02330138 -0.03313913 1.097517 0.0005621738 0.04495136
3 -0.07402715  0.05795748 -0.08425153 1.088317 0.0036218179 0.04314380
4 -0.02503318  0.01942089 -0.02877018 1.094085 0.0004237595 0.04175295
5  0.10356793 -0.07468192  0.12904835 1.061979 0.0084330151 0.03417165
6  0.08685842 -0.05753566  0.11879955 1.057559 0.0071483665 0.02969158
```

Influence Measures

The `influence.measures()` function in R provides a comprehensive set of influence diagnostics, including:

- **DFBETAS**: Measures the effect of deleting each observation on the estimated coefficients.
- **DFFITS**: Measures the effect of deleting each observation on the fitted values.
- **COVRATIO**: Measures the effect of deleting each observation on the covariance matrix of the estimated coefficients.

- **Cook's Distance**: Measures the influence of deleting each observation on the estimated regression coefficients.
- **Hat values (Leverage)**: Measures the influence of each observation on the fitted values.

Interpretation

- **DFBETAS**: Values close to 0 indicate little influence. Large absolute values (e.g., > 1 for small datasets) suggest influential observations.
- **DFFITS**: Values larger than 1 or $\frac{2\sqrt{k+1}}{n}$ (where k is the number of predictors and n number of observations) suggest influential observations.
- **COVRATIO**: Values far from 1 indicate influential observations affecting the covariance matrix.
- **Cook's Distance**: Values larger than 1 indicate highly influential observations.
- **Hat values (Leverage)**: Values larger than $\frac{2k}{n}$ suggest high leverage points.

Important Considerations:

- **Context Matters:** Cutoff values are guidelines, and the appropriate cutoff can vary depending on the specific dataset and research question.
- **Don't Automatically Remove:** Don't remove influential points solely based on these diagnostics. Investigate why they are influential and whether they represent genuine data points or potential errors.
- **Consider Alternatives:** If influential points are problematic, consider using robust regression techniques or transforming the data.
- **Report Findings:** Always report any influential points identified in your analysis.

Studentized Residuals

Studentized residuals are residuals divided by an estimate of their standard deviation. They help identify outliers and are calculated as:

$$r_i = \frac{\epsilon_i}{s(\epsilon_i)}$$

, where where ϵ_i is the residual for observation i and $s(\epsilon_i)$ is its standard deviation.

Interpretation

If $|r_i| > 2$, then there is potential outliers.

```
head(
 matrix(
  rstudent(chap6data1.model1),
  ncol=1)
 ) #Studentized Residuals
```

```
           [,1]
[1,] -0.5434298
[2,] -0.1527504
[3,] -0.3967728
[4,] -0.1378281
[5,]  0.6860716
[6,]  0.6791302
```

Use the following code to add the diagnostic measures and the fitted values to the original data set.

```
within(chap6data1, {
FittedValues <- fitted(chap6data1.model1)
  Residuals <- residuals(chap6data1.model1)
  StudentizedResiduals <- rstudent(chap6data1.model1)
  HatValues <- hatvalues(chap6data1.model1)
  CooksDistance <- cooks.distance(chap6data1.model1)
  ObsNumber <- 1:nrow(chap6data1)
}) |> head()
```

```
  BodyWeight MetabolicRate ObsNumber CooksDistance  HatValues
1       49.9          1079         1  0.0073475621 0.04664371
2       50.8          1146         2  0.0005621738 0.04495136
3       51.8          1115         3  0.0036218179 0.04314380
4       52.6          1161         4  0.0004237595 0.04175295
5       57.6          1325         5  0.0084330151 0.03417165
6       61.4          1351         6  0.0071483665 0.02969158
  StudentizedResiduals Residuals FittedValues
1           -0.5434298 -84.49711     1163.497
2           -0.1527504 -23.85069     1169.851
3           -0.3967728 -61.91022     1176.910
4           -0.1378281 -21.55784     1182.558
5            0.6860716 107.14452     1217.855
6            0.6791302 106.31832     1244.682
```

Creating Influence.Cutoffs function

```r
influence.cutoffs <- function(model){
  p <- length(model$coefficients)
  n <- length(model$residuals)

  DFBETAS <- 2/(n^0.5)
  DFFITS <- 2*(p/n)^0.5
  COVRATIO.lwr <- 1-3*p/n
  COVRATIO.upr <- 1+3*p/n
  cook.D <- 4/n
  HATDIAG <- 2*p/n

  list(DFBETAS=DFBETAS,DFFITS=DFFITS,COVRATIO.lwr=COVRATIO.lwr,
       COVRATIO.upr=COVRATIO.upr,cook.D=cook.D,HATDIAG=HATDIAG)
}

influence.cutoffs(chap6data1.model1)
```

```
$DFBETAS
[1] 0.3015113

$DFFITS
[1] 0.4264014

$COVRATIO.lwr
[1] 0.8636364

$COVRATIO.upr
[1] 1.136364

$cook.D
[1] 0.09090909

$HATDIAG
[1] 0.09090909
```

#Assess assumptions: normality and homoscedasticity

```
shapiro.test(chap6data1.model1$residuals)
```

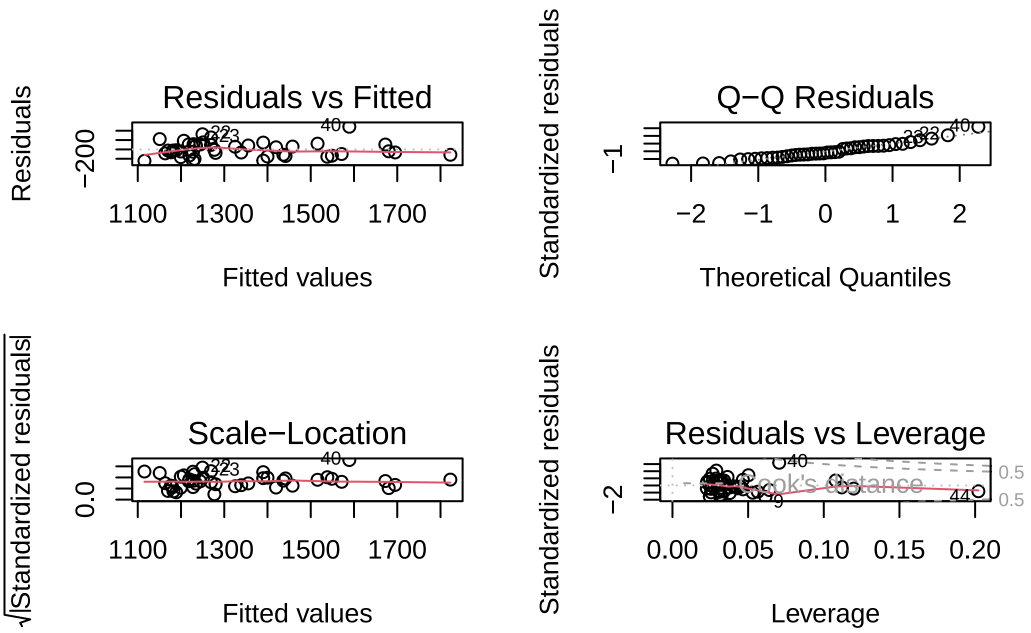
```

    Shapiro-Wilk normality test

data:  chap6data1.model1$residuals
W = 0.95657, p-value = 0.09681
```

Since **p > 0.05**, you **fail to reject the null hypothesis**. That means there is **no strong evidence** that the residuals deviate from normality. So, **normality assumption holds** reasonably well for your model.

```
par(mfrow=c(2,2))
base::plot(chap6data1.model1)
```



```
par(mfrow=c(1,1))
```

Prediction

```r
chap6data1.pred <- data.frame(BodyWeight=c(70,100))
predict(chap6data1.model1,newdata=chap6data1.pred,interval="confidence")
```

```
       fit      lwr      upr
1 1305.394 1256.398 1354.389
2 1517.179 1448.157 1586.202
```

```r
predict(chap6data1.model1,newdata=chap6data1.pred,interval="prediction")
```

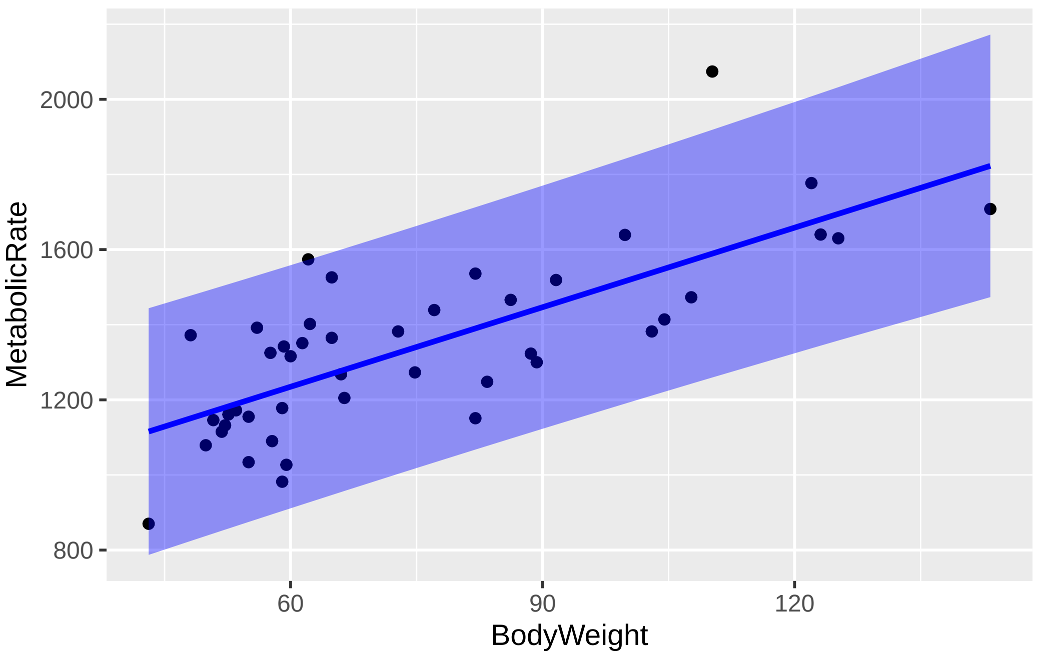
```
       fit       lwr      upr
1 1305.394  982.9833 1627.804
2 1517.179 1191.1245 1843.234
```

```r
chap6data1.predictions <- cbind(chap6data1,
                                predict(chap6data1.model1,
                                        newdata=chap6data1,
                                        interval="prediction")
                                )
head(chap6data1.predictions)
```

```
  BodyWeight MetabolicRate      fit      lwr      upr
1       49.9          1079 1163.497 837.4843 1489.510
2       50.8          1146 1169.851 844.1015 1495.600
3       51.8          1115 1176.910 851.4429 1502.378
4       52.6          1161 1182.558 857.3076 1507.808
5       57.6          1325 1217.855 893.7909 1541.920
6       61.4          1351 1244.682 921.3198 1568.044
```

```r
ggplot(chap6data1.predictions,aes(x=BodyWeight,y=MetabolicRate)) +
 geom_point() +
 geom_line(aes(y = fit),colour = "blue", size = 1) +
 geom_ribbon(aes(ymin = lwr, ymax = upr),  fill = "blue", alpha = 0.4)
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



IV- Exercises

Exercise 1

With the data from C&S Example 7.1, p. 64 (chap6exer1.csv): Perform a simple linear regression of Area (Y) on Stem length (X) and give the equation of the best-fitting regression line, as well as the coefficient of determination ($R2$). Interpret the $R2$ value. Also test the assumptions underlying this regression model, and clearly state your conclusions with regard to the validity of the fitted model.

Exercise 2

With the data from C&S Example 7.2, p. 69 (chap6exer2.csv): Perform a simple linear regression of wheat yield (Y) on nitrogen fertilizer (X) and give the equation of the best-fitting regression line. Predict the wheat yield for a nitrogen fertilizer level of $X = 80$kg nitrogen per hectare (also include the confidence and prediction intervals in your answer, and clearly interpret).

Exercise 3

Do C&S Exercise 7.2 (p. 84). The data is in the file chap6exer3.csv. A study was undertaken to find out if tree diameter measurements 1.5m above ground level can be used to predict heights for a certain species. The measurements of 12 trees are shown in the csv file. Complete a thorough regression analysis (significance, R-squared, assumptions).