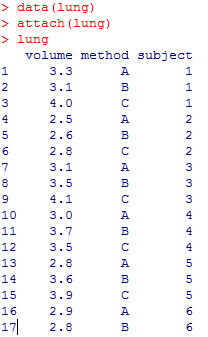
Lung Data Report by Juan Sunez

**Problem 1: load the lung dataset from the ISwR package**



**The lung data frame has 18 rows and 3 columns. It contains data on three different methods of determining human lung volume.**

**Variables:**

**volume : a numeric vector, measured lung volume.**

**method : a factor with levels A, B, and C.**

**subject : a factor with levels 1–6.**

**Develop R commands (script file) to determine if there are any differences in mean lung volume between the method or subjects. Make sure to first describe the data numerically and graphically, before running the appropriate analyses. Check any necessary assumptions and write a thorough conclusion of the results (Word document)**

library(ISwR)

data(lung)

attach(lung)

lung

boxplot(volume~as.factor(method), col=c("red","green","blue"), xlab="Method", ylab="Volume", main="Volume by Method")

win.graph()

boxplot(volume~subject, col=c(2:7), xlab="Subject", ylab="Volume", main="Volume by Subject")

summary(aov(volume ~ as.factor(method) + subject))

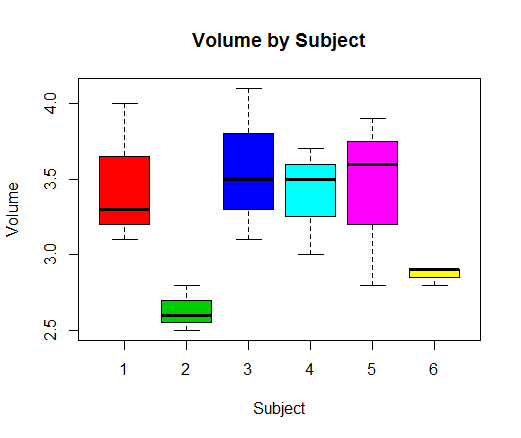
TukeyHSD(aov(volume ~ as.factor(method) + subject))

plot(resid(aov(volume ~ as.factor(method) + subject)))

hist(resid(lm(volume ~ as.factor(method) + subject)))

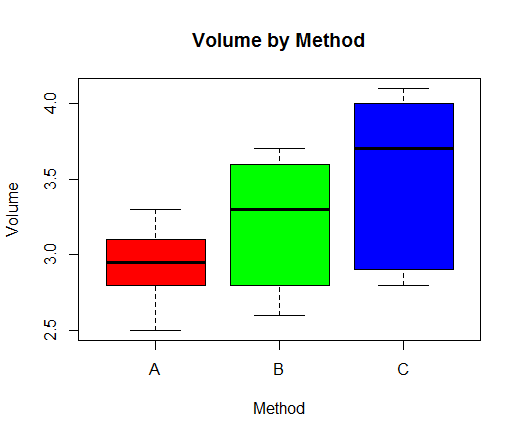
shapiro.test(resid(lm(volume ~ as.factor(method) + subject)))

*Figure 1.1 shows the values of volume by subject. As can be seen, there is not much difference between the mean of the volume of Subject 1, 3, 4 and 5 which are 3.47, 3.57, 3.4 and 3.43 respectively. However, that groups differ significantly from the group made up of the remaining variables, Subject 2 and 6 which have a volume mean of 2.93 and 2.87 respectively.*



**Figure 1.1**

*Figure 1.2 shows the values of volume by method. There seems to be a significant difference between the mean values of volume for each of the methods.*



**Figure 1.2**

*Figure 1.3 shows the ANOVA table which reveal the pr(>F) value for each variable to be below 0.05 indicating that both the method and subject variables are significant when explaining the variation in volume.*

**ANOVA**

|  |
| --- |
| Df Sum Sq Mean Sq F value Pr(>F) |
| as.factor(method) 2 1.0811 0.5406 6.495 0.0156 \* |
| subject 5 2.1828 0.4366 5.246 0.0127 \* |
| Residuals 10 0.8322 0.0832 |

**Figure 1.3**

*Figure 1.4 shows the results of the TukeyHSD which identifies from where the variation in those variables is coming from. As it could have been deduced by carefully analyzing the boxplots above, the results in Figure 1.4 show that there is a significant difference between method A and C while B was not significantly different than neither A or C. In addition, as it may have also been deduced by the boxplots, there is a significant different between subject 2 and 6 versus the rest of the subjects 1, 3, 4 and 5. However, statistically speaking, while there is no significant statistical difference between the mean of subject 2 and 6, there is a significant difference between subject 2 and the other subjects and not between subject 6 and the other subjects. As can be seen in the TukeyHSD table, the p-adj values of subject 2 compared to subjects 1, 3, 4 and 5 are below 0.05 while the values for subject 6 are all above 0.05.*

**TukeyHSD**

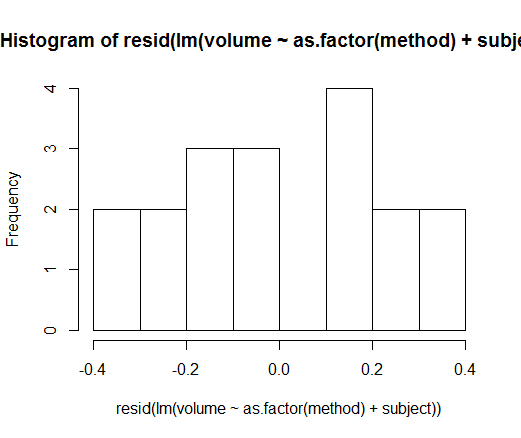
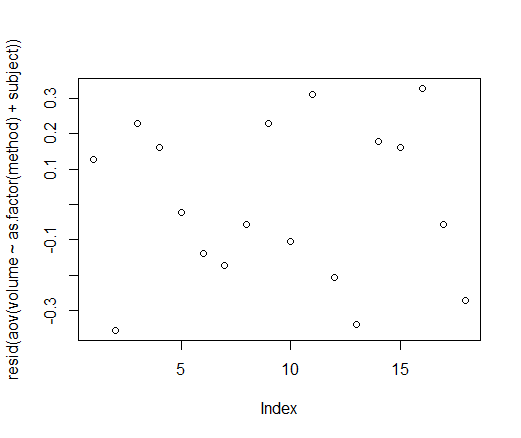
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| $`as.factor(method)` | | |  |  |  |  |
| diff lwr upr p adj | | | | | |  |
| B-A 0.2833333 -0.1732445 0.7399112 0.2520218 | | | | | |  |
| C-A 0.6000000 0.1434222 1.0565778 0.0122174 | | | | | |  |
| C-B 0.3166667 -0.1399112 0.7732445 0.1885516 | | | | | |  |
|  |  |  |  |  |  |  |
| $subject | |  |  |  |  |  |
| diff lwr upr p-adj | | | | | | |
| 2-1 -0.83333333 -1.65145666 -0.0152100 0.0453208 | | | | | | |
| 3-1 0.10000000 -0.71812333 0.9181233 0.9976751 | | | | | | |
| 4-1 -0.06666667 -0.88479000 0.7514567 0.9996680 | | | | | | |
| 5-1 -0.03333333 -0.85145666 0.7847900 0.9999891 | | | | | | |
| 6-1 -0.60000000 -1.41812333 0.2181233 0.1976352 | | | | | | |
| 3-2 0.93333333 0.11521000 1.7514567 0.0237972 | | | | | | |
| 4-2 0.76666667 -0.05145666 1.5847900 0.0696726 | | | | | | |
| 5-2 0.80000000 -0.01812333 1.6181233 0.0562068 | | | | | | |
| 6-2 0.23333333 -0.58479000 1.0514567 0.9108816 | | | | | | |
| 4-3 -0.16666667 -0.98479000 0.6514567 0.9766629 | | | | | | |
| 5-3 -0.13333333 -0.95145666 0.6847900 0.9912276 | | | | | | |
| 6-3 -0.70000000 -1.51812333 0.1181233 0.1066365 | | | | | | |
| 5-4 0.03333333 -0.78479000 0.8514567 0.9999891 | | | | | | |
| 6-4 -0.53333333 -1.35145666 0.2847900 0.2905205 | | | | | | |
| 6-5 -0.56666667 -1.38479000 0.2514567 0.2404375 |  |  |  |  |  |  |

**Figure 1.4**

**Assumption Check:**

*The graphics below show that the residual are normally distributed satisfying the normality assumption.*

**Residual**



*In addition, the p-value greater than 0.05 in the Shapiro-Wilk test shown below also indicate that the residual satisfy the normality assumption.*

***Shapiro-Wilk normality test***

*data: resid(lm(volume ~ as.factor(method) + subject))*

*W = 0.93947, p-value = 0.2838***Problem 2: A government researcher is analyzing the relationship between retail sales and the gross national product (GNP). He also wonders whether there are significant differences in retail sales related to the quarters of the year, compared to the baseline of Q4. He collects ten years of quarterly data.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **Quarter** | **Retail sales (in millions)** | **GNP (in billions)** |
| 2002 | Q1 | 696048 | 9740.5 |
| 2002 | Q2 | 753211 | 9983.5 |
| 2002 | Q3 | 746875 | 10048.0 |
| 2002 | Q4 | 792622 | 10184.9 |
| 2003 | Q1 | 704757 | 10206.2 |
| 2003 | Q2 | 779011 | 10350.9 |
| 2003 | Q3 | 756128 | 10332.2 |
| 2003 | Q4 | 827829 | 10463.1 |
| 2004 | Q1 | 717302 | 10549.7 |
| 2004 | Q2 | 790486 | 10634.7 |
| 2004 | Q3 | 792657 | 10749.1 |
| 2004 | Q4 | 833877 | 10832.2 |
| 2005 | Q1 | 741233 | 10940.2 |
| 2005 | Q2 | 819940 | 11073.6 |
| 2005 | Q3 | 831222 | 11321.2 |
| 2005 | Q4 | 875437 | 11508.3 |
| 2006 | Q1 | 795916 | 11707.8 |
| 2006 | Q2 | 871970 | 11864.2 |
| 2006 | Q3 | 873695 | 12047.3 |
| 2006 | Q4 | 938213 | 12216.6 |
| 2007 | Q1 | 836952 | 12486.3 |
| 2007 | Q2 | 932713 | 12613.0 |
| 2007 | Q3 | 940880 | 12848.7 |
| 2007 | Q4 | 987085 | 12994.1 |
| 2008 | Q1 | 897180 | 13264.0 |
| 2008 | Q2 | 987406 | 13423.3 |
| 2008 | Q3 | 978211 | 13514.8 |
| 2008 | Q4 | 1018775 | 13683.2 |
| 2009 | Q1 | 923997 | 13859.8 |
| 2009 | Q2 | 1016136 | 14087.6 |
| 2009 | Q3 | 1002312 | 14302.9 |
| 2009 | Q4 | 1062803 | 14489.9 |
| 2010 | Q1 | 953358 | 14520.7 |
| 2010 | Q2 | 1032919 | 14647.3 |
| 2010 | Q3 | 1006551 | 14689.2 |
| 2010 | Q4 | 966329 | 14317.2 |
| 2011 | Q1 | 839625 | 14172.2 |
| 2011 | Q2 | 919646 | 14164.2 |
| 2011 | Q3 | 926265 | 14281.9 |
| 2011 | Q4 | 985649 | 14442.8 |

1. **Load this data however you wish into R.**

*retail.gdp.data <- read.csv("C:/Users/jsunez/Desktop/DSS 665/Week 4/HW Week 4/Retail\_GDP\_Data.csv", header = T)*

*attach(retail.gdp.data)*

> head(retail.gdp.data)

Year Quarter Retail GNP

2002 Q1 696048 9740.5

2002 Q2 753211 9983.5

2002 Q3 746875 10048.0

2002 Q4 792622 10184.9

2003 Q1 704757 10206.2

2003 Q2 779011 10350.9

1. **Which variable is the response ?**



*The response variable is Retail.*

1. **Since the researcher is interested in whether or not the Quarter of the year has an impact on retail sales, write R commands to create the dummy variables you need to recode the “Quarter” column, using Q4 as baseline.**

> q1.rc <- c(rep(0, length(Quarter)))

> for(i in 1:length(Quarter)){

+ if(Quarter[i] == "Q1"){

+ q1.rc[i] = 1

+ }

+ }

> q1.rc

[1] 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0

> q2.rc <- c(rep(0, length(Quarter)))

> for(i in 1:length(Quarter)){

+ if(Quarter[i] == "Q2"){

+ q2.rc[i] = 1

+ }

+ }

> q2.rc

[1] 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0

> q3.rc <- c(rep(0, length(Quarter)))

> for(i in 1:length(Quarter)){

+ if(Quarter[i] == "Q3"){

+ q3.rc[i] = 1

+ }

+ }

> q3.rc

[1] 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0

> retail.gnp.recoded <- cbind(Retail, GNP, q1.rc, q2.rc, q3.rc)

> head(retail.gnp.recoded)

Retail GNP q1.rc q2.rc q3.rc

[1,] 696048 9740.5 1 0 0

[2,] 753211 9983.5 0 1 0

[3,] 746875 10048.0 0 0 1

[4,] 792622 10184.9 0 0 0

[5,] 704757 10206.2 1 0 0

[6,] 779011 10350.9 0 1 0

1. **Using R, find the best MLR model for the data (disregard Year variable). Is the overall model significant? Justify with ANOVA hypotheses and p-value. Are the individual variables significant? If not, remove the insignificant variables stepwise, until all remaining variables are significant**

*The overall model is significant as indicated by the Adjucted R-squared of 0.9127 and p-value < 2.2e-16. From the result of the linear model we can see that all the variable have a level of significance. GNP, Q1 and Q3 have the greater level of significance when in determining the variation in retail sales with pr(>|t|) values well under 0.05 than Q2 whose pr(>|t|) = 0.06088. Although Q2 is significant to a 0.1 confidence level, it is removed from the model in the stepwise regression process. Once Q2 is removed, the remaining variables, GNP, Q1 and Q3, are all statistically significant factor in determining the retail sales. From it we create the model:*

*Retail = 247385.333 + 53.410(GNP) - 85400.416(Q1) – 24913.720(Q3)*

1. **Predict retail sales in quarter 1 if GNP equals $13,000,000,000,000 (13,000 billion)**

*Retail Sales = 247385.333 + 53.410(13,000) - 85400.416(1) – 24913.720(0)*

*Retail Sales = 856,314.917 millions*

1. **Interpret the meaning of all significant slope coefficients in the “reduced” model.**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 247385.333 38110.365 6.491 1.54e-07 \*\*\*

GNP 53.410 3.021 17.681 < 2e-16 \*\*\*

q1.rc -85400.416 12233.952 -6.981 3.48e-08 \*\*\*

q3.rc -24913.720 12209.953 -2.040 0.0487 \*

* *(Intercept) of 247385.333 is the y-intercept of the line equation.*
* *GNP coefficient of 53.41 means that for every unit of GNP, retail sales will increase by a factor of 53.41 from the y-intercept.*
* *Q1 coefficient of -85400.416 means that effect of Q1 decreases the value of retail sales by -85400.416.*
* *Q3 coefficient of -24913.72 means that effect of Q3 decreases the value of retail sales by -24913.72.*

1. **Look at the residual plots. Are there any patterns? Should we worry about the assumption of MLR?**

|  |  |
| --- | --- |
|  |  |

*Both the scatterplot and histogram of the residuals show a concerning pattern. The scatterplot does not show a normal distribution as also supported by the histogram of the residual which shows the data is skewed to the left. The Shapiro-Wilk test of the residuals validates our concern by regarding Normality. As shown below, the p-value of the Shapiro-Wilk test is less than 0.05 failing to satisfy the Normality test.*

> shapiro.test(resid.retail.gnp)

Shapiro-Wilk normality test

data: resid.retail.gnp

W = 0.92949, p-value = 0.0155