Multivariate statistics – ANOVA and linear regression

Suckwon Hong

Ulsan National Institute of Science and Technology, 50 UNIST-gil Ulsan, Republic of Korea
amoeba94@unist.ac.kr

Question 1. Run a two-way ANOVA with the advertising dataset. Factor 1 is day of week (Monday – Friday). Factor 2 is section of newspaper (news, business, sports).

```
            Df Sum Sq Mean Sq F value   Pr(>F)
Day          4 146.83   36.71  20.910 8.52e-10 ***
Section      2  53.73   26.87  15.304 8.50e-06 ***
Day:Section  8 135.77   16.97   9.667 1.12e-07 ***
Residuals   45  79.00    1.76
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1. Two-way ANOVA results

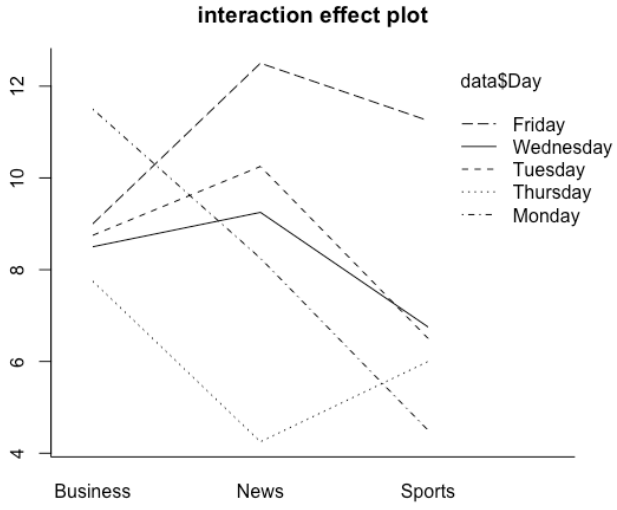
1.1. What are the hypotheses?

Answer: there are three null hypotheses as follows: 1) There is no difference in the means of *Factor 1*, 2) There is no difference in the means of *Factor 2,* 3) There is no interaction between factors.

1.2. Interpret the main effects and interaction effect.

Answer: It is found that there exist main effects of variable *Day* and *Section* on the dependent variable, *Responses.* Additionally, it is also found that there was a statistically significant interaction between the effects of *Day* and *Section* on *Response.*

1.3. Draw a plot of the interaction effect



Question 2. Run a simple linear regression with the demand dataset. X is price difference and Y is demand.

```
Call:
lm(formula = demand.y ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45713 -0.21121 -0.04898  0.14314  0.84961

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.81409    0.07988   97.82  < 2e-16 ***
price.difference.x 2.66521    0.25850   10.31 4.88e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3166 on 28 degrees of freedom
Multiple R-squared:  0.7915,    Adjusted R-squared:  0.7841
F-statistic: 106.3 on 1 and 28 DF,  p-value: 4.881e-11
```

Figure 3. Regression results

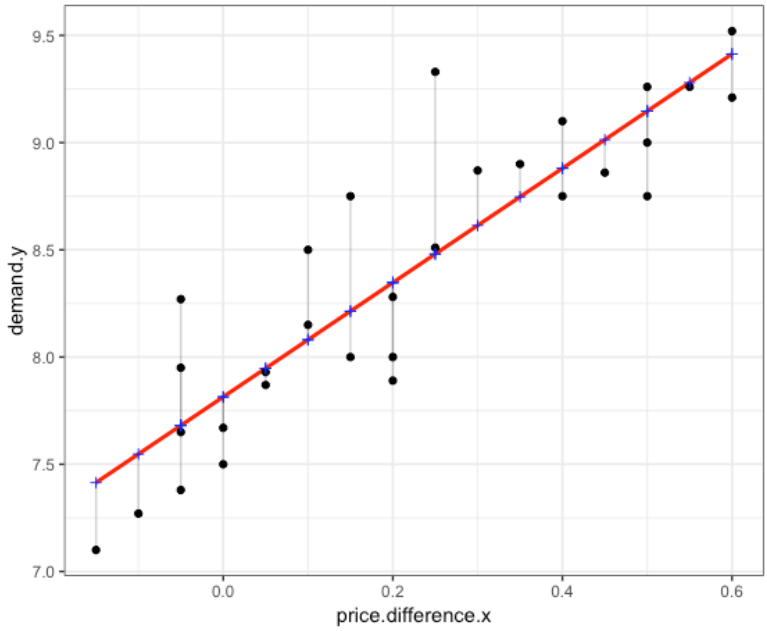


Figure 4. Data points plot with regression line

2.1. What is the R^2? Interpret R^2.

Answer: The value of R^2, coefficient of determination, is equal to 0.7915. It indicates that independent variables explain about 79% of the total variance of dependent variable.

2.2. Interpret the *F*-statistics.

Answer: F-statistics test the null hypothesis that 'the fitness of intercept-only model and the full model is same.' In this analysis, since the value of F-statistics is smaller than the critical value, we reject the null hypothesis.

2.3. Interpret the parameter estimates (constant and slope).

Answer: The value of constant and slope are 7.8141 and 2.6652, respectively. From the value of constant, we can expect that an average *demand* of 7.8141 with no *price difference.* In addition, the slope indicates that for every additional unit of *price difference,* we can expect *demand* to increase by average of 2.6652.

2.4. What is the regression equation?

Answer:

$$(demand)_i = \beta_1 * (price\ difference)_i + \beta_0 + \epsilon_i$$

2.5. What is the predicted value for x=0?

Answer: By plugging in the value of 0 to *price difference*, we get the predicted value of 7.8141 for the new value.

$$7.8141 = \beta_1 * 0 + \beta_0$$

Code.

Prerequisite: R programming language is applied for the assignment. And, we convert file extension of the data from .xlsx to .csv.

Question 1.

```r
### Multivaraite statistics assignment 1 - question 1

## data load
data <- read.csv("advertisement.csv",header = TRUE)

## ANOVA without interaction
res.aov2 <- aov(Response ~ Day + Section, data = data)
summary(res.aov2)

## ANOVA with interaction
res.aov3 <- aov(Response ~ Day*Section, data=data)
summary(res.aov3)

## interaction plot
par(mar=c(3,3,3,3))
interaction.plot(data$Section, data$Day, data$Response, bty='l', main="interaction effect plot")
```

Question 2.

```r
### Multivaraite statistics assignment 1 - question 2

## download package
install.packages("ggplot2") #for visualization
library(ggplot2)

## data load
data <- read.csv("demand.csv",header = TRUE)
data <- data[,2:3] #drop the first column

## make a regression model
lm.1 <- lm(demand.y ~ ., data = data) #model making
summary(lm.1) #model summary

## plotting
data$residulas <- resid(lm.1)
data$predicted <- predict(lm.1)

ggplot(data, aes(x = price.difference.x, y = demand.y)) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +  # Plot regression slope
  geom_segment(aes(xend = price.difference.x, yend = predicted), alpha = .2) +
  geom_point() +
  geom_point(aes(y = predicted), shape = 3, color = "blue") +
  theme_bw()  # Add theme for cleaner look

## predict the value of demand for new datum (i.e. 0)
new_x <- data.frame(price.difference.x=c(0))
predict_x0 <- predict(lm.1, newdata = new_x)
```