R Assignment:03

Keya Adhyaru

Question 1 - Local Advertising

Upon request, the local newspaper will provide data to help potential advertisers target their ads. The paper provides a random sample of 12 advertisements per day, 4 from each section of the paper, for the previous week and the number of inquiries the ad generated for the business.

a) Ignoring any potential interaction between Day and Section, are there any differences in the average number of inquiries by day?

Answer:

Considering the dataset given, we will be using ANOVA to analyse our data, with categorical predictors- Day and Section and numeric continuous response- Inquiry.

Setting up Hypothesis:

H_0 : There is no difference in the average number of inquiries by day.

H_A : Not all average number of inquiries per day are equal.

```
library(readr)
advertising <- read_csv("Path/to/advertising.csv")
```

```
## Rows: 60 Columns: 3
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): Day, Section
## dbl (1): Inquiries
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Applying additive model to avoid potential interaction of Section on Day.
advers_mod <- aov(Inquiries~ Day+ Section, data= advertising)
anova_result <- summary(advers_mod)
anova_result
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Day          4 146.83   36.71   9.059 1.19e-05 ***
## Section      2  53.73   26.87   6.630  0.00269 **
## Residuals   53 214.77    4.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the applied method, we can observe that p-val= 1.19e-05< 0.05, so we will reject the null hypothesis and further conclude that there exist significant effect of Day on average Inquiry received.

b) If "yes" to a), which day would you prefer to advertise on if maximizing inquiries was your objective?

Answer:

To further analyse the data, we will conduct post-hoc test, Tukey HSD Test. It allows for all possible pairwise comparisons while keeping the family-wise error rate low.

```
#FInding which day has more effect on larger number of Inquiries.
TukeyHSD(advers_mod, "Day")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Inquiries ~ Day + Section, data = advertising)
##
## $Day
##                           diff        lwr         upr     p adj
## Monday-Friday       -2.83333333 -5.1540343 -0.51263240 0.0094385
## Thursday-Friday     -4.91666667 -7.2373676 -2.59596574 0.0000019
## Tuesday-Friday      -2.41666667 -4.7373676 -0.09596574 0.0373566
## Wednesday-Friday    -2.75000000 -5.0707009 -0.42929907 0.0125928
## Thursday-Monday     -2.08333333 -4.4040343  0.23736760 0.0981402
## Tuesday-Monday       0.41666667 -1.9040343  2.73736760 0.9863219
## Wednesday-Monday     0.08333333 -2.2373676  2.40403426 0.9999753
## Tuesday-Thursday     2.50000000  0.1792991  4.82070093 0.0287660
## Wednesday-Thursday   2.16666667 -0.1540343  4.48736760 0.0780657
## Wednesday-Tuesday   -0.33333333 -2.6540343  1.98736760 0.9941367
```

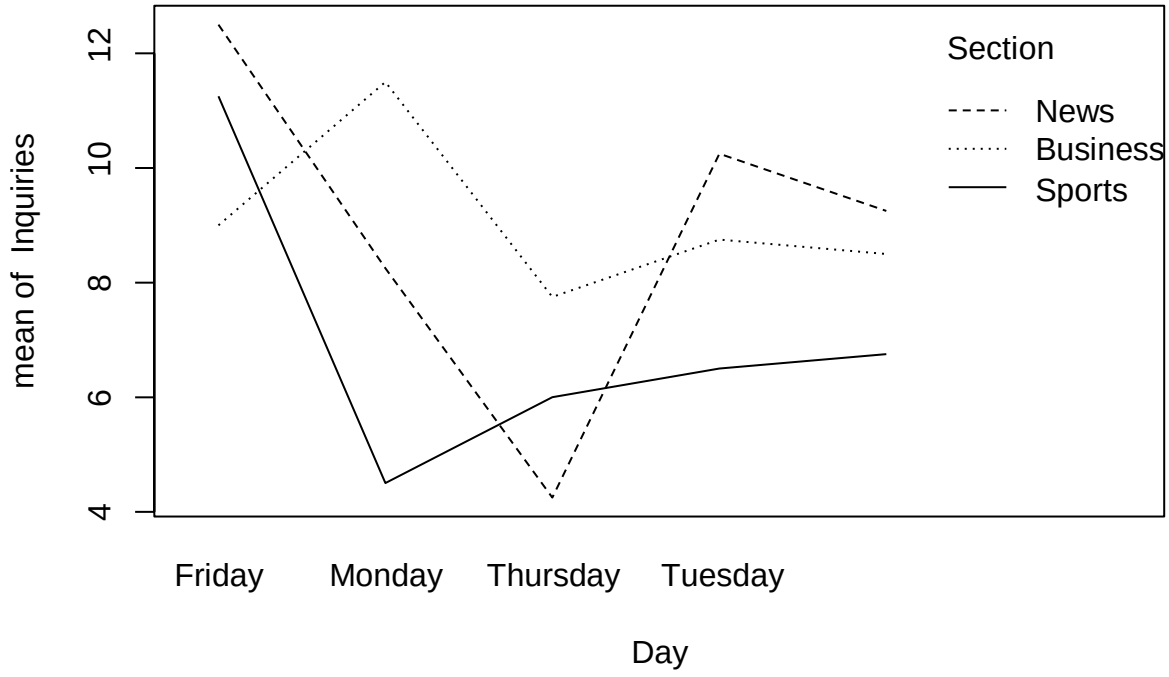
From the applied method, we can observe that the highest magnitude of difference amongst all the days examined from the result, is -4.91666667 for Thursday-Friday with the lowest adjusted p-value, 0.0000019<0.05. We hereby conclude that Friday is the day when highest number of Inquires were experienced.

c) Create an interaction plot for Day and Section on Inquiries. Based on this plot, does there appear to be an interaction; why or why not? (see the examples in the help file for interaction.plot on how to create these plots)

Answer:

Constructing interaction plot between day and section from the given dataset:

```
with(advertising, interaction.plot(x.factor = Day, trace.factor = Section, response = Inquiries))
```



From this plot, we can surely say that there exists an interaction between these two variables, as all the lines are depicting the sections are crossing each other at several different points. These fields are either converging or diverging at one or the other day which shows the evidence that there exists interaction.

d) Test for an interaction between Day and Section. Does your result have an effect on your answer in part a) of this question? **Bonus for determining which section and day combination sees the most inquiries using emmeans** [1 mark].

Answer:

To test the interaction between both the variables, we would apply multiplicative model on ANOVA.

```
mult_anova <- aov(Inquiries~ Day * Section, data= advertising)
mult_result <- summary(mult_anova)
mult_result
```

```
##             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
```

From the result, we can observe that p-value of Day=8.52e-10 <0.05, p-value of section = 8.50e-06<0.05 and p-value of interaction= 1.12e-07<0.05. Considering these low p-value we reject the null hypothesis and can conclude that there is strong significance of both individual variables as well as their interaction on the

response variable. Comparing the result of the test that we performed in part a), we witness larger magnitude of p-value, yet same conclusion- having significant effect on average number of inquiries experienced.

```
# determining which section and day combination sees the most inquiries using emmeans

library(emmeans)
emmeans(mult_anova, spec = ~ Day | Section)
```

```
## Section = Business:
##  Day       emmean    SE df lower.CL upper.CL
##  Friday      9.00 0.662 45     7.67    10.33
##  Monday     11.50 0.662 45    10.17    12.83
##  Thursday    7.75 0.662 45     6.42     9.08
##  Tuesday     8.75 0.662 45     7.42    10.08
##  Wednesday   8.50 0.662 45     7.17     9.83
##
## Section = News:
##  Day       emmean    SE df lower.CL upper.CL
##  Friday     12.50 0.662 45    11.17    13.83
##  Monday      8.25 0.662 45     6.92     9.58
##  Thursday    4.25 0.662 45     2.92     5.58
##  Tuesday    10.25 0.662 45     8.92    11.58
##  Wednesday   9.25 0.662 45     7.92    10.58
##
## Section = Sports:
##  Day       emmean    SE df lower.CL upper.CL
##  Friday     11.25 0.662 45     9.92    12.58
##  Monday      4.50 0.662 45     3.17     5.83
##  Thursday    6.00 0.662 45     4.67     7.33
##  Tuesday     6.50 0.662 45     5.17     7.83
##  Wednesday   6.75 0.662 45     5.42     8.08
##
## Confidence level used: 0.95
```

By approaching emmean , we have came to conclusion that highest number of inquiries were reported on friday for section news.

e) The newspaper currently charges the same amount per ad for any day and any section in the newspaper. Assuming you work for the paper, do you have any recommendations for the paper in terms of pricing?

Answer:

If I were to work for the paper company, after analyzing the interaction plot and reading the conclusion of part d), considering the average inquiries for particular day per section, I would charge more for Friday for News and sports articles, Monday for Business articles, while put the medium charge on thursday specific to Business articles , tuesday for News and Business articles. Lastly, I would plan to put minimal charges on advertisements for the sports section for monday, tuesday, wednesday and thursday.

Question 2 - Ground Water

Water treatment plants add bicarbonate to water in order to keep microorganisms in the system happy and healthy. The data in groundwater.csv is a sample of pH is measured on a logarithmic scale from 0 to 14 and bicarbonate levels are measured in parts per million (ppm)

Parametric

a) Use a linear model and parametric method to determine if there is a relationship between bicarbonate levels and pH in the water. (Still check and comment on assumptions, but use a parametric method regardless!)

Answer:

Following are the assumptions to be met in order to conduct parametric test for linear model:

- **Independence of the observations:** Each subject should belong to only one group. There is no relationship between the observations in each group. Having repeated measures for the same participants is not allowed.
- **Outliers:** There must be no significant outliers in any cell of the design.
- **Normality:** the data for each design cell should be approximately normally distributed.
- **Homogeneity of variances:** The variance of the outcome variable should be equal in every cell of the design.

We assume that all observations are independent in nature as they are randomly sampled and there no outliers in the observation.

```
library(readr)
groundwater <- read_csv("Path/to/groundwater.csv")
```

```
## Rows: 34 Columns: 2
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (2): pH, Bicarbonate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Checking normality:

```
shapiro.test(groundwater$Bicarbonate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  groundwater$Bicarbonate
## W = 0.97692, p-value = 0.6736
```

From the shapiro test of normality we got, p-val= 0.6736 >0.05 for Bicarbonate and p-value = 0.4981 >0.05 for pH.We can conclude that the observations of the dataset shows normality.

Now, we will build a simple linear regression model for the given data . Firstly, we will check if these variables have a linear relationship amongst them by conducting a correlation test.

```
cor(groundwater$pH, groundwater$Bicarbonate)
```

```
## [1] -0.3395105
```

From the above test, we get the pearson's coefficient r= -0.3395105. It suggests that pH and bicarbonates possess weak negative linear relationships.

Despite of being weak relationship, we will evaluate the model,

```
linear_mod <- lm(formula = pH ~ Bicarbonate, data = groundwater)
summary(linear_mod)
```

```
##
## Call:
## lm(formula = pH ~ Bicarbonate, data = groundwater)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04049 -0.41037 -0.01841  0.24995  1.15107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.097595   0.228714  35.405   <2e-16 ***
## Bicarbonate -0.003052   0.001495  -2.042   0.0495 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.479 on 32 degrees of freedom
## Multiple R-squared:  0.1153, Adjusted R-squared:  0.08762
## F-statistic: 4.169 on 1 and 32 DF,  p-value: 0.04948
```

Linear model constructed can be written as follows:

$$\hat{□} = 8.097595 - 0.003052□□□□□$$

Setting up hypothesis:

H_0 : □1=0 (There is no linear relationship between pH and Bicarbonate)

H_A : □1 ≠ 0 (There is linear relationship between pH and Bicarbonate)

Results:

While conducting the test, We got p-val:0.04948 < 0.05 . So, we reject the null hypothesis and further conclude that there is enough evidence to prove that there exists a linear relationship among pH and Bicarbonate.

b) Plot the data and regression line. Include appropriate labels and title. Bonus for including confidence and prediction intervals on this plot [2 marks].

Answer:

Plotting the scatterplot and regression line for the respective data,

```r
#adding prediction and store it in the datasetname "new_data"
pred.int <- predict(linear_mod, interval="prediction")
```

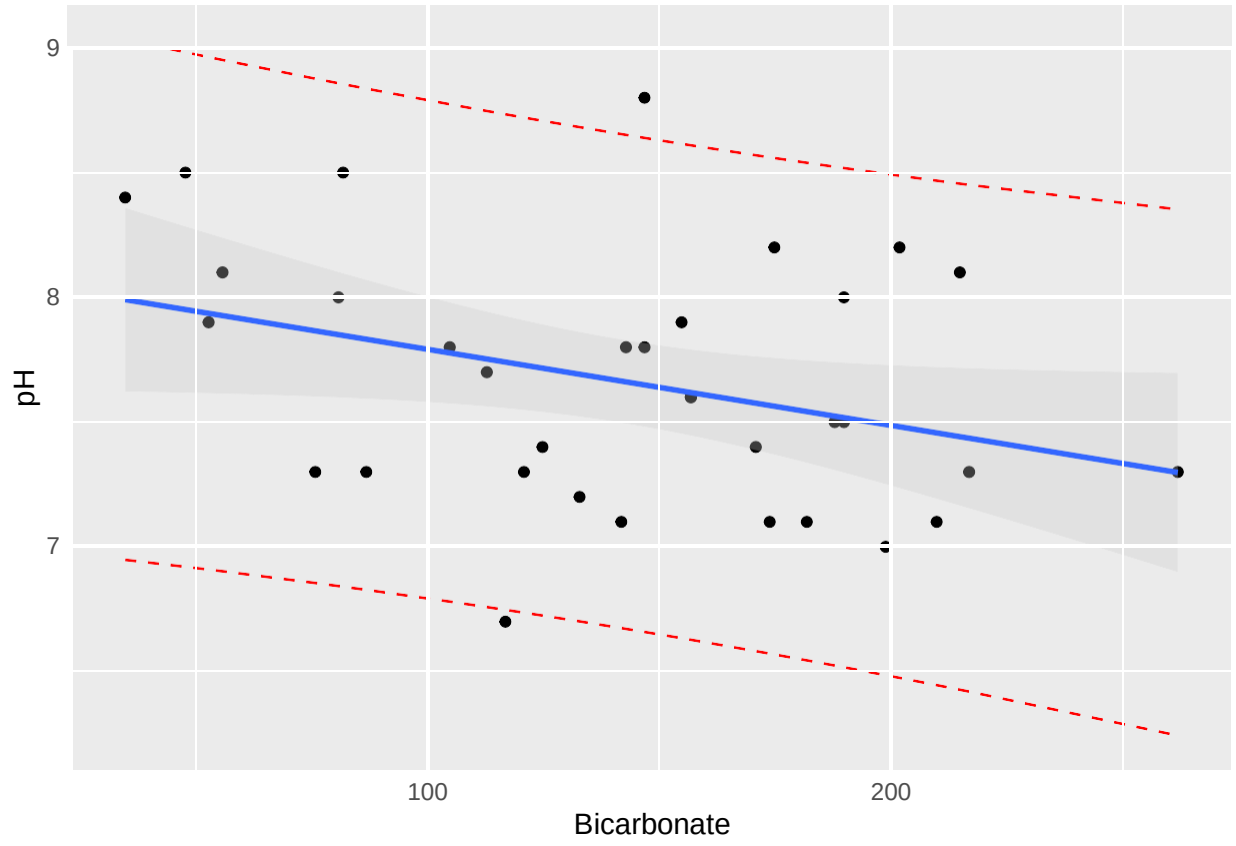
```
## Warning in predict.lm(linear_mod, interval = "prediction"): predictions on current data refer to _fu
```

```r
new_data <- cbind(groundwater, pred.int)

#Regression line + 95% confidence intervals
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```r
p <- ggplot(new_data, aes(Bicarbonate, pH)) + geom_point() + stat_smooth(method= lm)

# Adding prediction intervals
p+geom_line(aes(y= lwr), color= "red", linetype= "dashed") + geom_line(aes(y=upr),color
= "red", linetype = "dashed")
```

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



Prediction interval with respect to given data:

```
pred.int
```

```
##         fit      lwr      upr
## 1  7.618406 6.627430 8.609382
## 2  7.566519 6.571939 8.561100
## 3  7.563467 6.568591 8.558344
## 4  7.523789 6.524235 8.523343
## 5  7.575676 6.581926 8.569426
## 6  7.661136 6.671104 8.651168
## 7  7.435277 6.419790 8.450763
## 8  7.517685 6.517274 8.518096
## 9  7.664188 6.674154 8.654223
## 10 7.517685 6.517274 8.518096
## 11 7.441381 6.427232 8.455530
## 12 7.490215 6.485501 8.494930
## 13 7.297929 6.243461 8.352398
## 14 7.777118 6.780421 8.773816
## 15 7.728284 6.736031 8.720537
## 16 7.850370 6.842618 8.858122
## 17 7.847318 6.840130 8.854506
## 18 7.456642 6.445683 8.467600
## 19 7.481059 6.474748 8.487369
## 20 7.624510 6.633781 8.615240
## 21 7.618406 6.627430 8.609382
## 22 7.648928 6.658813 8.639042
## 23 7.691658 6.701177 8.682139
## 24 7.935831 6.908740 8.962921
## 25 7.926674 6.901978 8.951371
## 26 7.752701 6.758522 8.746881
## 27 7.990770 6.947751 9.033788
## 28 7.716075 6.724562 8.707588
## 29 7.865631 6.854926 8.876336
## 30 7.951091 6.919844 8.982339
## 31 7.648928 6.658813 8.639042
## 32 7.740492 6.747350 8.733635
## 33 7.542102 6.544899 8.539305
## 34 7.832057 6.827556 8.836559
```

Confidence interval with respect to given data:

```
#predict confidence interval for the linear model
conf.int <- predict(linear_mod, interval="confidence")
conf.int
```

```
##         fit      lwr      upr
## 1  7.618406 7.445560 7.791252
## 2  7.566519 7.374080 7.758958
## 3  7.563467 7.369507 7.757428
## 4  7.523789 7.307109 7.740470
## 5  7.575676 7.387579 7.763773
## 6  7.661136 7.493789 7.828483
## 7  7.435277 7.154109 7.716444
## 8  7.517685 7.297086 7.738284
```

```
## 9  7.664188 7.496825 7.831552
## 10 7.517685 7.297086 7.738284
## 11 7.441381 7.165083 7.717679
## 12 7.490215 7.250857 7.729574
## 13 7.297929 6.898248 7.697611
## 14 7.777118 7.574024 7.980213
## 15 7.728284 7.548261 7.908307
## 16 7.850370 7.598565 8.102175
## 17 7.847318 7.597780 8.096856
## 18 7.456642 7.192297 7.720986
## 19 7.481059 7.235087 7.727031
## 20 7.624510 7.453087 7.795933
## 21 7.618406 7.445560 7.791252
## 22 7.648928 7.481092 7.816763
## 23 7.691658 7.521676 7.861640
## 24 7.935831 7.615275 8.256387
## 25 7.926674 7.613873 8.239476
## 26 7.752701 7.562348 7.943054
## 27 7.990770 7.622355 8.359184
## 28 7.716075 7.540177 7.891973
## 29 7.865631 7.602256 8.129006
## 30 7.951091 7.617456 8.284727
## 31 7.648928 7.481092 7.816763
## 32 7.740492 7.555633 7.925352
## 33 7.542102 7.336542 7.747662
## 34 7.832057 7.593594 8.070521
```

c) Provide a 95% parametric confidence interval for the true slope in the linear model.

Answer:

Constructing 95% confidence interval for the true slope in the linear model:

```
confint(linear_mod, parm = "Bicarbonate", level = 0.95)
```

```
##                      2.5 %       97.5 %
## Bicarbonate -0.006096981 -7.33803e-06
```

The 95% confidence interval of the true slope is (-0.006097, -7.33803e-06)

Non-Parametric

d) Use a linear model and non-parametric method to determine if there is a relationship between bicarbonate levels and pH in the water.

Answer:

If our assumptions are not satisfied, we can use non-parametric methods to run the tests.

Setting hypothesis:

H_0 : $\beta_1 = 0$ (There is no linear relationship between pH and Bicarbonate)

H_A : $\beta_1 \neq 0$ (There is linear relationship between pH and Bicarbonate)

```r
B <- 10000
grdwater_sim <- groundwater[, c("pH", "Bicarbonate")]
grdwatr1_sim <- rep(NA, B)

for (i in 1:B){
  idx_sim <- sample(1:nrow(groundwater), size = nrow(groundwater), replace = FALSE)
  grdwater_sim$pH <- grdwater_sim$pH[idx_sim]
  mod_sim <- lm(pH ~ Bicarbonate, data = grdwater_sim)
  grdwatr1_sim[i] <- mod_sim$coefficients[2]
}

non_param_p_val <- (length(which(abs(grdwatr1_sim) >=
                        abs(linear_mod$coefficients[2]))) + 1) / (B + 1)
non_param_p_val
```

```
## [1] 0.04849515
```

By conducting non parametric approach, we got the p-value = 0.0484952< 0.05 so we reject the null hypothesis and further conclude that there is evidence of linear relationship among Bicarbonate and pH.

e) Provide a 95% non-parametric confidence interval for the true slope parameter in the linear model.

Answer:

Providing non-parametric confidence interval for the true slope parameter in the linear model:

```r
library(boot)

lm_wrapper <- function(x, index){
  mod <- lm(pH ~ Bicarbonate, data = x[index, ])
  mod$coefficients[2]
}

b1_bs <- boot(groundwater, statistic = lm_wrapper, R = 2000)
boot.ci(b1_bs, type = "bca")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = b1_bs, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   (-0.0055,  0.0000 )
## Calculations and Intervals on Original Scale
```

We are 95% confident that true slope lies under the interval mentioned under Bca parameter.

(Note: I am unable to mention the result value in conclusion statement as everytime while knitting this document it generates different results due to bootstrapping).