Problem Set 10

Prateek Kumar
December 02, 2018

Table of Contents

Problem 1: Categorical Predictors

On each of day of one week, we sampled 100 random company stocks and examined their trading price. Each day a different set of stocks was sampled at random from the NYSE and NASDAQ published prices.

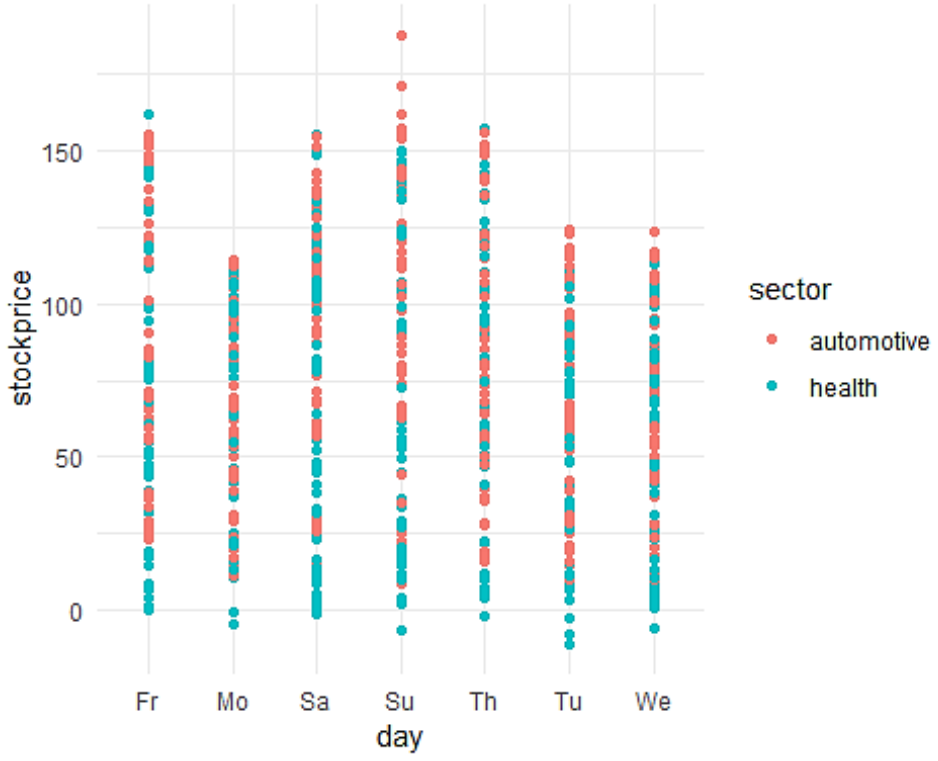
```
library(ggplot2)
data <- read.csv("ps10data.csv")

head(data)

##   day     sector stockprice
## 1  Mo automotive      55.11
## 2  Mo automotive      85.10
## 3  Mo     health      99.67
## 4  Mo automotive      19.79
## 5  Mo automotive      69.68
## 6  Mo automotive      61.97
```

This is stored in a matrix. For a regression or ANOVA, we really need each one

```
ggplot(data,aes(x=day,y=stockprice)) + geom_point(aes(color=sector)) +theme_minimal()
```



For this problem, we want to determine, using a number of methods, which days differed from which other days. In each case, run the test, and answer the question in 1-2 sentences describing what you found. Use a p=.05 as a criterion for determining whether an effect isstatistically significant.

1. First use a contrast that will compare each day to Monday, and report which of the days had prices significantly higher than Monday

```
day.0 <- c("Mo","Tu","We","Th","Fr","Sa","Su")
data$day<- factor(data$day,levels=day.0) #add the level to months variable
aggregate(data$stockprice,list(data$day),mean)
```

```
##   Group.1       x
## 1      Mo 60.5618
## 2      Tu 59.5182
## 3      We 60.2386
## 4      Th 80.7623
## 5      Fr 81.5663
## 6      Sa 78.4041
## 7      Su 83.7771
```

```
model1 <- lm(stockprice~day, data=data)
model1
```

```
##
## Call:
## lm(formula = stockprice ~ day, data = data)
##
## Coefficients:
## (Intercept)        dayTu        dayWe        dayTh        dayFr
##     60.5618      -1.0436      -0.3232      20.2005      21.0045
##       daySa        daySu
##     17.8423      23.2153
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = stockprice ~ day, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.497 -33.876  -0.053  36.118 103.973
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.5618     4.2292  14.320  < 2e-16 ***
## dayTu        -1.0436     5.9809  -0.174 0.861533
## dayWe        -0.3232     5.9809  -0.054 0.956920
## dayTh        20.2005     5.9809   3.377 0.000772 ***
## dayFr        21.0045     5.9809   3.512 0.000474 ***
## daySa        17.8423     5.9809   2.983 0.002953 **
## daySu        23.2153     5.9809   3.882 0.000114 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.29 on 693 degrees of freedom
```

```
## Multiple R-squared:  0.05869,    Adjusted R-squared:  0.05054
## F-statistic: 7.202 on 6 and 693 DF,  p-value: 1.808e-07
```

We can see from the above results that the for days Thursday, Friday, Saturday and Sunday the prices are significantly higher than Monday.

The model estimates each level (in this case days in a week) with its actual mean. When we do the summary of the model we see that for Tuesday (-1.0236) and Wednesday (-0.3232) the values are almost same as that of Monday but the values are subsequently higher for other days.

2. Then, use successive difference coding of the day variable to determine which days of the week differed significantly from the previous day.

```
library(MASS)
contrasts(data$day)<-contr.sdif(levels(data$day))
model2 <- lm(stockprice~day, data=data)
model2

##
## Call:
## lm(formula = stockprice ~ day, data = data)
##
## Coefficients:
## (Intercept)     dayTu-Mo     dayWe-Tu     dayTh-We     dayFr-Th
##     72.1183      -1.0436       0.7204      20.5237       0.8040
##    daySa-Fr     daySu-Sa
##     -3.1622       5.3730

summary(model2)

##
## Call:
## lm(formula = stockprice ~ day, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.497 -33.876  -0.053  36.118 103.973
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  72.1183     1.5985  45.117  < 2e-16 ***
## dayTu-Mo     -1.0436     5.9809  -0.174 0.861533
## dayWe-Tu      0.7204     5.9809   0.120 0.904162
## dayTh-We     20.5237     5.9809   3.432 0.000636 ***
## dayFr-Th      0.8040     5.9809   0.134 0.893104
## daySa-Fr     -3.1622     5.9809  -0.529 0.597174
## daySu-Sa      5.3730     5.9809   0.898 0.369309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.29 on 693 degrees of freedom
```

```
## Multiple R-squared:  0.05869,    Adjusted R-squared:  0.05054
## F-statistic: 7.202 on 6 and 693 DF,  p-value: 1.808e-07
```

Further on using successive difference coding of the day variable we can see that Thursday differs with a significant amount viz. 23.5237 with respect to Wednesday.

3. Use pairwise.t.test function to compute all pairwise t-tests and the holm correction between days of the week. Describe concisely which days differed from which other days.

```
pairwise.t.test(data$stockprice, data$day)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data$stockprice and data$day
##
##    Mo     Tu     We     Th     Fr     Sa
## Tu 1.0000 -      -      -      -      -
## We 1.0000 1.0000 -      -      -      -
## Th 0.0100 0.0066 0.0089 -      -      -
## Fr 0.0071 0.0044 0.0066 1.0000 -      -
## Sa 0.0295 0.0199 0.0272 1.0000 1.0000 -
## Su 0.0022 0.0012 0.0018 1.0000 1.0000 1.0000
##
## P value adjustment method: holm
```

On using pairwise.t.test() function to compute all pairwise t-tests and the holm correction between days of the week we see that we have either value 1.0 or any other value. If the value is 1.0, it means there is no difference but any other value means there is a difference between the days.

Thus, Monday, Tuesday and Wednesday does not differ from each other but differ from other four days and vice versa for Thursday, Friday, Saturday and Sunday which does not differ from each other but differs from Monday, Tuesday and Wednesday.

4. Use an aov() model to predict stock price by day, and then compute Tukey HSD test on all pairwise comparisons using the Tukey test. Does the result differ from part 3?

```
TukeyHSD(aov(stockprice~day,data=data))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = stockprice ~ day, data = data)
##
## $day
##          diff        lwr      upr     p adj
## Tu-Mo -1.0436 -18.729386 16.64219 0.9999976
```

```
## We-Mo -0.3232 -18.008986 17.36259 1.0000000
## Th-Mo 20.2005   2.514714 37.88629 0.0135627
## Fr-Mo 21.0045   3.318714 38.69029 0.0085579
## Sa-Mo 17.8423   0.156514 35.52809 0.0463876
## Su-Mo 23.2153   5.529514 40.90109 0.0021786
## We-Tu  0.7204 -16.965386 18.40619 0.9999997
## Th-Tu 21.2441   3.558314 38.92989 0.0074315
## Fr-Tu 22.0481   4.362314 39.73389 0.0045691
## Sa-Tu 18.8859   1.200114 36.57169 0.0275391
## Su-Tu 24.2589   6.573114 41.94469 0.0010864
## Th-We 20.5237   2.837914 38.20949 0.0112984
## Fr-We 21.3277   3.641914 39.01349 0.0070715
## Sa-We 18.1655   0.479714 35.85129 0.0396268
## Su-We 23.5385   5.852714 41.22429 0.0017622
## Fr-Th  0.8040 -16.881786 18.48979 0.9999995
## Sa-Th -2.3582 -20.043986 15.32759 0.9997075
## Su-Th  3.0148 -14.670986 20.70059 0.9988009
## Sa-Fr -3.1622 -20.847986 14.52359 0.9984289
## Su-Fr  2.2108 -15.474986 19.89659 0.9997990
## Su-Sa  5.3730 -12.312786 23.05879 0.9727959
```

On using an aov() model to predict stock price by day, and further then computing Tukey HSD test on all pairwise comparisons using the Tukey test which tries to adjust the p-values we observed that the results obtained here are almost same as that obtained in Q3.

The p-values for comparison between Monday, Tuesday and Wednesday is more than 0.05 and is almost 1.0 and for remaining four days it's less than 0.05 stating difference between the days and vice versa for Thursday, Friday, Saturday and Sunday.

5. Compute a kruskall-wallis test to see if the non-parametric test shows stock price depended on day-of-week.

```
kruskal.test(stockprice~day,data=data)

##
##  Kruskal-Wallis rank sum test
##
## data:  stockprice by day
## Kruskal-Wallis chi-squared = 36.113, df = 6, p-value = 2.621e-06

#summary(kruskal.test(stockprice~day,data=data))
```

On computing a kruskall-wallis test we see that the chi-squared value is 36.113 which is greater than the degrees of freedom value (6) thus it rejects the NULL hypothesis moreover the p-value is less than 0.05 which further confirms our result. Henceforth, the stock price depends on day-of-week.

6. Compute a one-way BayesFactor ANOVA and report the Bayes factor score determining if day-of-week impacted stock price.

```
library(BayesFactor)

bfmodel <- anovaBF(stockprice~day,data=data)
bfmodel

## Bayes factor analysis
## --------------
## [1] day : 41865.66 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

On computing a one-way BayesFactor ANOVA we got a bayes factor score of 41865.66 stating very strong evidence for alternate hypothesis. This confirms that day-of-week definitely impacts the stock price.

2. Multi-way ANOVA and regression

The stocks were sampled from two different sectors (health and automotive). Was there a difference in outcome based on sector? What about when day day-of-week is considered. Report a standard (Type-I) ANOVA F-test for:

1. The effect of sector on its own (a one-way test), and

```
lm1 <- lm(stockprice~sector, data=data)
summary(lm1)

##
## Call:
## lm(formula = stockprice ~ sector, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.583 -36.600  -2.367  34.652 108.738
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    79.012      2.320  34.055  < 2e-16 ***
## sectorhealth  -13.479      3.244  -4.155 3.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.91 on 698 degrees of freedom
## Multiple R-squared:  0.02413,    Adjusted R-squared:  0.02274
## F-statistic: 17.26 on 1 and 698 DF,  p-value: 3.659e-05

anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: stockprice
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## sector      1   31780   31780  17.263 3.659e-05 ***
## Residuals 698 1284995    1841
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#summary(aov(lm(stockprice~sector, data=data)))

## oneway.test(stockprice~sector, var.equal=T, data = data) #checks difference in mean

##    One-way analysis of means
##
## data:  stockprice and sector
## F = 17.263, num df = 1, denom df = 698, p-value = 3.659e-05
```

Here we see that the p-value is less than 0.05 this interprets that different sectors have different average stock price and thus sector has a significant impact towards stock price.

2. whether sector has an effect *after* day-of-week is considered:

```
anova(lm(stockprice~day+sector, data=data))

## Analysis of Variance Table
##
## Response: stockprice
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## day         6   77286   12881  7.3853 1.128e-07 ***
## sector      1   32536   32536 18.6545 1.796e-05 ***
## Residuals 692 1206952    1744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We still have the p-value less than 0.05 stating that both day and sector have an impact on stock price.

3. whether the results differ if sector is included in the model first

```
anova(lm(stockprice~sector+day, data=data))

## Analysis of Variance Table
##
## Response: stockprice
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## sector      1   31780   31780 18.2209 2.242e-05 ***
## day         6   78043   13007  7.4576 9.360e-08 ***
## Residuals 692 1206952    1744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here as well we have the p-value less than 0.05 stating that both day and sector have an impact on stock price.

When we compare the above models we see that there is a slight difference in the sum-squared deviations when sector is considered after the day-of-week else in the other two cases the value is same viz. 31780 because the predictors are orthogonal and balanced. But the F-ratio is different because the F-ratio is the ratio between the mean square deviation (which is same in both models) and the residuals (which change).

In our more complex model, the residuals go down because we are explaining them through another reliable predictor.

I would prefer the third model to test the effect because:

- ***From model 2 and model 3 we already identified that both sector and day-of-week have an impact on the stock so that rejects using model 1.***
- ***And amongst model 2 and 3 we see that model 3 is most close to model 1 i.e. the sum-squared deviations is same in the case of sector.***

Thus, concentrating on sector model 3 is the best fit.