HW-4

Dennis Goldenberg

2024-02-15

Homework 4 - Predictive Modeling in Finance and Insurance

```r
library(ggplot2)
library(readxl)
#library(dplyr)
```

1. ANOVA, one factor

```r
#read in data, remove empty column
phosData <- read_excel("Table 6.19-PlasmaPhosphate-1.xls", skip = 2,
                       sheet = "Sheet1", .name_repair = "unique_quiet")
phosData <- phosData[1:12, 1:3]
colnames(phosData) <- c("H_O","N_O", "C")
```

a. Test of difference of means

Our hypotheses are as follows:

$

$$H_0 : \mu_{H_O} = \mu_{N_O} = \mu_C \text{ and } H_a : \text{ at least one different}$$

$

To test, I first calculate the means $\bar{x}_{H_O}, \bar{x}_{N_O}, \bar{x}_C$. There are 3 levels, so 2 degrees of freedom between sum of squares, and $n - 3$ degrees of freedom in the error sum of squares. I calculate the mean sum of squares for both and generate the F-statistic:

```r
means <- colMeans(phosData, na.rm = TRUE)
oMean <- sum(phosData, na.rm = TRUE)/sum(!is.na(phosData))
MSE <- 0
MSB <- 0
for (i in colnames(phosData)) {
  MSE <- MSE + sum((phosData[,i] - means[i])^2, na.rm = TRUE)
  MSB <- MSB + sum(!is.na(phosData[,i]))*((means[i] - oMean)^2)
}
MSE <- MSE/(sum(!is.na(phosData)) - (dim(phosData)[2]))
MSB <- unname(MSB/(dim(phosData)[2] - 1))
sprintf("Test statistic: %f", MSB/MSE)
```

```
## [1] "Test statistic: 11.650806"
```

I then calculate the p-value of this statistic, using $ndf = 2$ and $ddf = n - 3$ for the test statistic:

```r
p_1a <- pf(MSB/MSE, df1 = dim(phosData)[2] - 1,
   df2 = sum(!is.na(phosData)) - (dim(phosData)[2]), lower.tail = F)
sprintf("p value: %f", p_1a)
```

```
## [1] "p value: 0.000208"
```

Note that $p_{1a} < 0.05$, meaning we have **enough evidence to reject** H_0; there is evidence that the means for different treatments provide different results.

1b. 95% confidence interval, difference of means

Normality is assumed; therefore, given that the estimate of $\mu_{H-O} - \mu_{N-O} = \bar{x}_{H-O} - \bar{x_{N-O}}$, the bounds for the confidence interval are: $

$$\bar{x}_{H_O} - \bar{x}_{N_O} \pm z_{.975}\hat{\text{SE}}(\bar{x}_{H_O} - \bar{x_{N_O}}) = \bar{x}_{H_O} - \bar{x_{N_O}} \pm 1.96 * S_P\sqrt{\frac{1}{n_{H_O}} + \frac{1}{n_{N_O}}}$$

$ Here, the pooled sample variance is $S_P = \sqrt{\frac{(n_{H_O}-1)s^2_{H_O} + (n_{N_O}-1)s^2_{N_O}}{n_{H_O}+n_{N_O}-2}}$. I thus calculate the lower and upper bound:

```r
n_HO <- sum(!is.na(phosData[,"H_O"]))
n_NO <- sum(!is.na(phosData[,"N_O"]))
S_P <- sqrt(((n_HO - 1)*var(phosData$H_O, na.rm = TRUE)
             + (n_NO - 1)*var(phosData$N_O, na.rm = TRUE))/
              (n_HO + n_NO - 2))
LCI <- unname(means["H_O"] - means["N_O"])[1] - 1.96*S_P*sqrt(1/n_HO + 1/n_NO)
UCI <- unname(means["H_O"] - means["N_O"])[1] + 1.96*S_P*sqrt(1/n_HO + 1/n_NO)
diffMeansconf <- c(LCI, UCI)
names(diffMeansconf) <- c("Lower Bound", "Upper Bound")
diffMeansconf
```

```
## Lower Bound Upper Bound
## -0.09881829  1.11472738
```

1c. Standard Residual Plot

2. ANOVA, two factor (unblanaced)

First, I read in the data and turn the explanatory variables into factors:

```
facData <- read_excel("Table 6.21-Unbalanced data-1.xls", skip = 2,
                      sheet = "Sheet1", .name_repair = "unique_quiet")
facData$factorA <- factor(facData$factorA)
facData$factorB <- factor(facData$factorB)
colnames(facData)
```

```
## [1] "factorA" "factorB" "data"
```

2a. Testing for interaction effects

I run the linear model that includes interactions:

```
lm2_inter <- lm("data ~ factorA + factorB + factorA*factorB", data = facData)
lm2_noInter <- lm("data ~ factorA + factorB", data = facData)
summary(lm2_noInter)
```

```
##
## Call:
## lm(formula = "data ~ factorA + factorB", data = facData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7857 -0.6964 -0.1786  0.3393  1.3571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.71429    0.72804   6.475 0.000644 ***
## factorAA2    0.07143    0.77610   0.092 0.929666
## factorAA3    3.00000    0.82134   3.653 0.010674 *
## factorBB2   -1.07143    0.65854  -1.627 0.154866
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.006 on 6 degrees of freedom
## Multiple R-squared:  0.7665, Adjusted R-squared:  0.6497
## F-statistic: 6.565 on 3 and 6 DF,  p-value: 0.02529
```

3. One-factor ANOVA

4. National Life Expectancies