

# HW-4

Dennis Goldenberg

2024-02-15

## Homework 4 - Predictive Modeling in Finance and Insurance

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

### 1. ANOVA, one factor

```
#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_0", "N_0", "C")
```

#### a. Test of difference of means

Our hypotheses are as follows:

$$H_0 : \mu_{H_0} = \mu_{N_0} = \mu_C \text{ and } H_a : \text{at least one different}$$

To test, I first calculate the means  $\bar{x}_{H_0}, \bar{x}_{N_0}, \bar{x}_C$ . 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:

```
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:

```
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:

$$\bar{x}_{H-O} - \bar{x}_{N-O} \pm z_{.975} \hat{SE}(\bar{x}_{H-O} - \bar{x}_{N-O}) = \bar{x}_{H-O} - \bar{x}_{N-O} \pm 1.96 * S_P \sqrt{\frac{1}{n_{H-O}} + \frac{1}{n_{N-O}}}$$

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

```
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
```

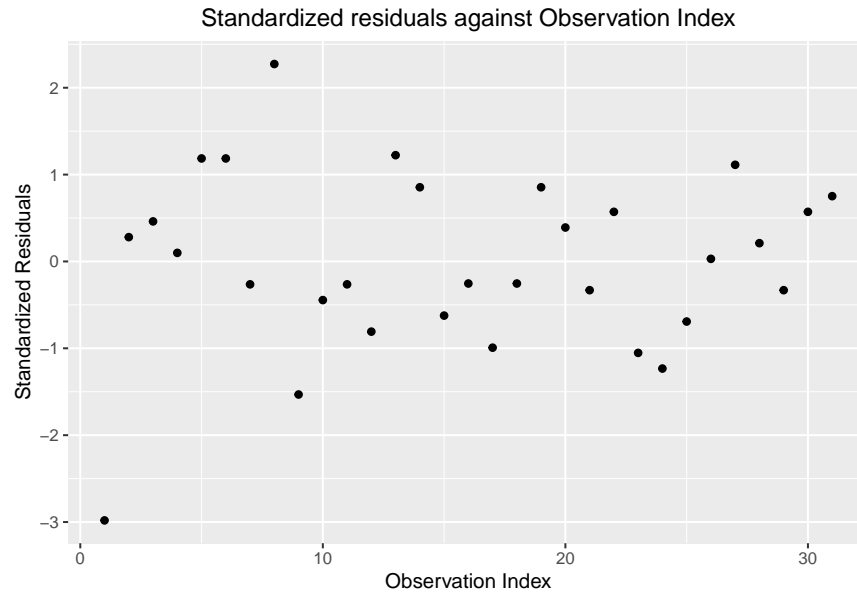
### 1c. Standard Residual Plot

I reread the data into a different format, time one column for group and one column for phosphate count for simplicity in implementation to the model:

```
phosDataC <- phosData <- read_excel("Table 6.19-PlasmaPhosphate-1.xls",
  skip = 2, sheet = "Sheet1",
  range = cell_cols("E:F"), .name_repair = "unique_quiet")
phosDataC$Group <- factor(phosDataC$Group)
lm_1c <- lm("phosphate ~ factor(Group)", data = phosDataC)
```

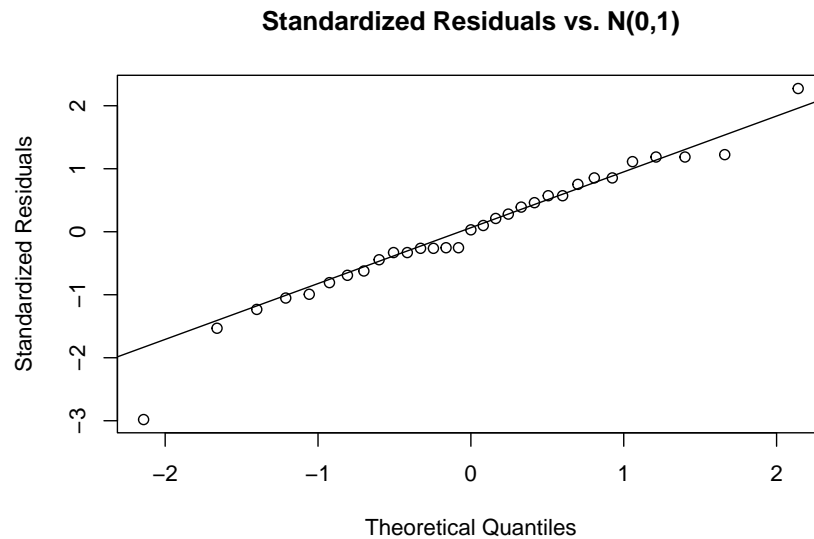
Then, I plot the standardized residuals against observation index:

```
std_resid <- rstandard(lm_1c)
resid_table <- data.frame(cbind(1:length(std_resid), std_resid))
colnames(resid_table) <- c("Index", "std_resid")
ggplot(data = resid_table) + geom_point(aes(x = Index, y = std_resid)) +
  ylab("Standardized Residuals") + xlab("Observation Index") +
  ggtitle("Standardized residuals against Observation Index") +
  theme(plot.title = element_text(hjust = 0.5))
```



Then, I produce a normal probability plot:

```
qqnorm(y = std_resid, ylab = "Standardized Residuals",
       xlab = "Theoretical Quantiles",
       main = "Standardized Residuals vs. N(0,1)")
qqline(y = std_resid)
```



Note that the fit is very good, with the exception of the 1st observation, having a standardized residual of almost  $-3$ . Thus, it can be concluded that the residuals are approximately normal.

## 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)
```

### 2a. Testing for interaction effects

Note that factor A has 3 levels, and factor B has 2 levels; therefore, there will be  $(3-1)(2-1) = 2$  coefficients corresponding to interaction effects in a full model. Assuming  $A_1$  and  $B_1$  as our reference groups, our hypotheses are:

$$H_0 : \beta_{AA_2BB_2} = \beta_{AA_3BB_2} = 0 \text{ (no interaction) and } H_a : \text{at least one } \beta_{AB} \neq 0$$

I run the linear model that includes interactions and the model without interactions:

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

These models are nested, and there are 2 coefficients being tested. Thus, in the  $F$  statistic, there are 2 degrees of freedom in the numerator and  $10 - 2 - (2 + 1 + 1) = 4$  in the denominator, I calculate the following F-statistic:

$$F_{stat} = \frac{SSE_{red} - SSE_{full}}{q * s_{full}^2}$$

```
f_2a <- (sum(residuals(lm2_noInter)^2) - sum(residuals(lm2_inter)^2))/
(2 * (summary(lm2_inter)$sigma)^2)
sprintf("F-stat = %f", f_2a)
```

```
## [1] "F-stat = 0.428571"
```

I calculate the p-value of this statistic:

```
sprintf("p-value = %f", pf(f_2a, df1 = 2, df2 = 4, lower.tail = FALSE))
```

```
## [1] "p-value = 0.678201"
```

Note that  $0.678201 > 0.05$ , so I **fail to reject**  $H_0$ .

### 2b. Testing effect of factor A, controlling for B

The hypotheses in this example are:

$$H_0 : \beta_{A_2} = \beta_{A_3} = 0 \text{ and } H_a : \text{at least one } \beta_A \neq 0$$

If the null hypothesis is true, than the level of factor A has no statistically significant bearing on the data. I have already generated the full model in this case (the one with no interactions). So, I generate the reduced model, and calculate the same f-statistic as in part 2a (here, there would be 2 degrees of freedom in the numerator, and  $10 - (3 + 1) = 6$  in the denominator as the full model has 3 parameters):

```
lm2_onlyB <- lm("data ~ factorB", data = facData)
f_2b <- (sum(residuals(lm2_onlyB)^2) - sum(residuals(lm2_noInter)^2))/
(2 * (summary(lm2_noInter)$sigma)^2)
sprintf("F-stat = %f", f_2b)
```

```
## [1] "F-stat = 9.023529"
```

I then calculate the p-value of this statistic:

```
sprintf("p-value = %f", pf(f_2b, df1 = 2, df2 = 6, lower.tail = FALSE))
```

```
## [1] "p-value = 0.015533"
```

Since  $0.155 < 0.05$ , I reject  $H_0$  at  $\alpha = 0.05$ ; factor A seems to be statistically significant.

### 2c. Testing the effect of factor A, standalone

In this case, the hypotheses are the same as in part B, but the full model is the model with only factor A in it, while the reduced model is a model with only an intercept (in this case the mean of the data). I calculate the F-statistic (in this case, the numerator has 2 degrees of freedom, as before, but denominator has  $10 - (2 + 1) = 7$  degrees of freedom, as the model with only factor A has no other parameters):

```
lm2_onlyA <- lm("data ~ factorA", data = facData)
lm2_justMean <- lm("data ~ factorB", data = facData)
f_2c <- (sum(residuals(lm2_justMean)^2) - sum(residuals(lm2_onlyA)^2))/
  (2 * (summary(lm2_onlyA)$sigma)^2)
sprintf("F-stat = %f", f_2c)
```

```
## [1] "F-stat = 6.233333"
```

I then calculate the p-value

```
sprintf("p-value = %f", pf(f_2c, df1 = 2, df2 = 7, lower.tail = FALSE))
```

```
## [1] "p-value = 0.027882"
```

Since  $0.0279 < 0.05$ , I reject  $H_0$  at  $\alpha = 0.05$ ; once again, factor A seems to be statistically significant.

### 2d. Comparing the results from 2b and 2c

The difference between our tests for significance in 2b and 2c is that the test in 2c measured the standalone predictive power of factor A on the data (there were no other predictors in the model) while the test in 2b controlled for Factor B; that is, the significance of the coefficients corresponding to different levels of A was a measure of the predictive power of factor A over the mean within a particular level of factor B. Note that the p-value for 2b was lower than 2c; this indicates that **factor A was more powerful in explaining the variance within each level of factor B** than over-all variance. This could be due to the unbalanced nature of the data or the fact that mean of data is different between factor B levels.

### 3. One-factor ANOVA

#### 3a. Calculating the sum sq for Within Groups

To calculate  $SSB$ , I first need to calculate the overall mean:

$$\bar{x} = \frac{n_C \bar{x}_C + n_{NO} \bar{x}_{NO} + n_{HO} \bar{x}_{HO}}{n_C + n_{NO} + n_{HO}} = \frac{12(2.93) + 10(3.58) + 11(4.19)}{11 + 10 + 12} = 3.547$$

I use the formula provided in the slides:

$$\begin{aligned} SSB &= \sum_{i \in C, NO, HO} n_i (\bar{x}_i - \bar{x})^2 \\ &= 12(2.93 - 3.547)^2 + 10(3.58 - 3.547)^2 + 11(4.19 - 3.547)^2 \\ &= \mathbf{9.1271} \end{aligned}$$

#### 3b. Filling in the other missing fields

The other values to be calculated are:

- Mean Squared Between ( $MSB$ )
- F statistic for between groups ( $F_B$ )
- p-value for f-stat above
- Sum of Squared Within ( $SSW$ )
- Degrees of Freedom Within ( $df_W$ )
- Mean Squared Within ( $MSW$ )

I start with  $MSB$ ,  $SSW$ ,  $df_W$  and  $MSW$ , using formulas from the slides:

$$\begin{aligned} MSB &= \frac{SSB}{df_B} = \frac{9.1271}{2} = \mathbf{4.56355} \\ SSW &= SST - SSB = 23.7218 - 9.1271 = \mathbf{14.5947} \\ df_W &= n_{\text{total}} - \text{Number of Groups} = 33 - 3 = \mathbf{30} \\ MSW &= \frac{SSW}{df_W} = \frac{14.5947}{30} = \mathbf{0.48649} \end{aligned}$$

Then, I calculate the F-statistic for between groups:

$$F_b = \frac{MSB}{MSW} = \frac{4.56355}{0.48649} = \mathbf{9.38056}$$

For the p-value, note that the statistic above comes from an F distribution with 2 numerator degrees of freedom and  $32 - 2 = 30$  denominator degrees of freedom:

```
sprintf("p-value = %f", pf(9.38056, df1 = 2, df2 = 30, lower.tail = FALSE))
```

```
## [1] "p-value = 0.000685"
```

#### 3c. Are the means the same?

The test statistic and p-value correspond to testing  $H_0 : \mu_C = \mu_{NO} = \mu_{HO}$ . Since  $0.000685 < 0.01$ , I would reject  $H_0$  at  $\alpha = .01, .05$ , and/or  $.1$ . There is statistically significant evidence that at least one mean is different.

#### 4. National Life Expectancies