STSCI 4520 Lab 1

Nick Gembs

1/25/2023

# Lab Exercise

Complete the exercise below in a SEPARATE R Markdown file, and turn it in as a pdf on canvas. DO NOT include the introduction to Rmarkdown from the top of this script.

Read in the “voice\_heights” data found in the github repo in datasets/voice\_heights.csv. The data contain the heights and voice categories for all people in a choir. Fit a linear model to the heights, using voice as a factor

dat <- read.csv("C:/Users/Nick/Documents/GitHub/statcomp2023/datasets/voice\_heights.csv")  
head(dat)

## gender voice height  
## 1 female soprano 64  
## 2 female soprano 62  
## 3 female soprano 66  
## 4 female soprano 65  
## 5 female soprano 60  
## 6 female soprano 61

h <- dat$height  
v <- as.factor(dat$voice)  
levels(v)

## [1] "alto" "bass" "soprano" "tenor"

m1 <- lm( h ~ v )  
summary(m1)

##   
## Call:  
## lm(formula = h ~ v)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1500 -1.8857 0.2821 1.2821 7.1143   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.8857 0.4250 152.655 < 2e-16 \*\*\*  
## vbass 5.8322 0.5855 9.961 < 2e-16 \*\*\*  
## vsoprano -0.6357 0.5969 -1.065 0.289   
## vtenor 4.2643 0.7049 6.050 1.54e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.515 on 126 degrees of freedom  
## Multiple R-squared: 0.5706, Adjusted R-squared: 0.5603   
## F-statistic: 55.8 on 3 and 126 DF, p-value: < 2.2e-16

Note that the regression output compares the height of each voice with the reference level.

Re-run the regression 3 more times, each time changing the reference level. Use the results to report the t-statistic for comparing the means of every pair of heights.

v <- as.factor(dat$voice)  
v <- relevel(v, ref = "bass")  
m1 <- lm( h ~ v )  
summary(m1)

##   
## Call:  
## lm(formula = h ~ v)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1500 -1.8857 0.2821 1.2821 7.1143   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 70.7179 0.4027 175.626 <2e-16 \*\*\*  
## valto -5.8322 0.5855 -9.961 <2e-16 \*\*\*  
## vsoprano -6.4679 0.5812 -11.129 <2e-16 \*\*\*  
## vtenor -1.5679 0.6916 -2.267 0.0251 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.515 on 126 degrees of freedom  
## Multiple R-squared: 0.5706, Adjusted R-squared: 0.5603   
## F-statistic: 55.8 on 3 and 126 DF, p-value: < 2.2e-16

*The regression above reports t-statistics for mean heights of choir group factor variables against a reference of bass. The t-statistic for alto vs bass is -9.961, soprano vs bass is -11.129, and tenor vs bass is -2.267*

v <- as.factor(dat$voice)  
v <- relevel(v, ref = "soprano")  
m1 <- lm( h ~ v )  
summary(m1)

##   
## Call:  
## lm(formula = h ~ v)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1500 -1.8857 0.2821 1.2821 7.1143   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.2500 0.4191 153.303 < 2e-16 \*\*\*  
## valto 0.6357 0.5969 1.065 0.289   
## vbass 6.4679 0.5812 11.129 < 2e-16 \*\*\*  
## vtenor 4.9000 0.7013 6.987 1.44e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.515 on 126 degrees of freedom  
## Multiple R-squared: 0.5706, Adjusted R-squared: 0.5603   
## F-statistic: 55.8 on 3 and 126 DF, p-value: < 2.2e-16

*The regression above reports t-statistics for mean heights of choir group factor variables against a reference of soprano. The t-statistic for alto vs soprano is 1.065, soprano vs bass is 11.129, and tenor vs soprano is 6.987*

v <- as.factor(dat$voice)  
v <- relevel(v, ref = "tenor")  
m1 <- lm( h ~ v )  
summary(m1)

##   
## Call:  
## lm(formula = h ~ v)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1500 -1.8857 0.2821 1.2821 7.1143   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.1500 0.5623 122.980 < 2e-16 \*\*\*  
## valto -4.2643 0.7049 -6.050 1.54e-08 \*\*\*  
## vbass 1.5679 0.6916 2.267 0.0251 \*   
## vsoprano -4.9000 0.7013 -6.987 1.44e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.515 on 126 degrees of freedom  
## Multiple R-squared: 0.5706, Adjusted R-squared: 0.5603   
## F-statistic: 55.8 on 3 and 126 DF, p-value: < 2.2e-16

*The regression above reports t-statistics for mean heights of choir group factor variables against a reference of tenor. The t-statistic for alto vs tenor is -6.050, tenor vs bass is 2.267, and tenor vs soprano is -6.987*

You can inspect the factor variables from an lm object using its xlevels field.

m1$xlevels

## $v  
## [1] "tenor" "alto" "bass" "soprano"

mat <- matrix(data = c(NA, 9.961,-1.065,6.050,-9.961,NA,-11.129,-2.267,1.065,11.129,NA,6.987,-6.050,2.267,-6.987,NA), ncol = 4, nrow = 4)  
v <- relevel(v, ref = "alto")  
colnames(mat) <- levels(v)  
rownames(mat) <- levels(v)  
mat

## alto tenor bass soprano  
## alto NA -9.961 1.065 -6.050  
## tenor 9.961 NA 11.129 2.267  
## bass -1.065 -11.129 NA -6.987  
## soprano 6.050 -2.267 6.987 NA