HW1 Solution

Gustavo

9/29/2022

Homework 1

Using the Gout data set:

1) Fit a linear model of the form `su~race+sex+age`, report your results, and summarize (in no more than three sentences) your conclusions.

```
DATA=read.table("https://raw.githubusercontent.com/gdlc/STAT_COMP/master/DATA/goutData.txt",header=TRU
fm1=lm(su~race+sex+age,data=DATA)
summary(fm1)
```

```
##
## Call:
## lm(formula = su ~ race + sex + age, data = DATA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4843 -0.9717 -0.1829  0.8276  5.4296
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.31975    0.81533   5.298 1.95e-07 ***
## raceW       -0.78212    0.16932  -4.619 5.22e-06 ***
## sexM         1.52853    0.14306  10.684  < 2e-16 ***
## age          0.02674    0.01299   2.058   0.0402 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.413 on 396 degrees of freedom
## Multiple R-squared:  0.2504, Adjusted R-squared:  0.2447
## F-statistic: 44.09 on 3 and 396 DF,  p-value: < 2.2e-16
```

Remarks: Both sex and race have significant effects on serum-urate levels. Male had higher SU than female, white had lower levels than black. Age has a marginally significant effect, with an indication of a slight increase of SU with age (0.0267 untis per year).

2) Consider now expanding the model to inclue race-by-sex interactions.

- Explain with words what an interaction term different than zero means in this model.
- Fit the model with the interaction term, report your results and conclusions.

```
fm2<-lm(su~race+race*sex+age,data=DATA)
summary(fm2)
```

```
##
```

```
## Call:
## lm(formula = su ~ race + race * sex + age, data = DATA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5293 -0.9190 -0.1923  0.8184  5.3810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.46104    0.82394   5.414 1.07e-07 ***
## raceW       -0.93555    0.21447  -4.362 1.65e-05 ***
## sexM         1.21196    0.30712   3.946 9.40e-05 ***
## age          0.02629    0.01299   2.024   0.0437 *
## raceW:sexM   0.40430    0.34712   1.165   0.2448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.413 on 395 degrees of freedom
## Multiple R-squared:  0.253,  Adjusted R-squared:  0.2454
## F-statistic: 33.44 on 4 and 395 DF,  p-value: < 2.2e-16
```

The estimated effect for the interaction was positive, suggesting that the difference between male and female is slightly higher for white than for black peopole. However, the evidence is weak, and the interaction is not significantly different than zero.

3) Consider now testing the hypothesis that sex has **any** effect on su (it could be an effect dependent on race or independent of it) versus the null that states that sex has no effect on su.

- Describe the null and the alternative hypothesis,
- Test the null using `anova()`, and
- Summarize your findings.

```
fm0=lm(su~race+age,data=DATA)
anova(fm0,fm2)
```

```
## Analysis of Variance Table
##
## Model 1: su ~ race + age
## Model 2: su ~ race + race * sex + age
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    397 1019.11
## 2    395  788.36  2    230.75 57.808 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remark: We conclude that sex has an effect on SU.

4) Reproducing the results of the F-test:

- Review the F-statistic in the class notes and
- Develop a function that takes as input two `lm` objects and return a table identical to the one produced by `anova()`.
- Test your function using the H0 and Ha you used in Q3.

```
myANOVA=function(fm0,fm1){
  df0=length(coef(fm0))
  df1=length(coef(fm1))
  n=length(residuals(fm1))
  RSS0=sum(residuals(fm0)^2)
  RSS1=sum(residuals(fm1)^2)

  FSTAT=((RSS0-RSS1)/(df1-df0))/(RSS1/(n-df1))

  ANS=cbind('RES.DF'=c(n-df0,n-df1),'RSS'=c(RSS0,RSS1),DF=c(NA,df1-df0),
            'Sum of Sq'=c(NA,RSS0-RSS1),
            'F'=c(NA,FSTAT),
            'pValue'=c(NA,pf(lower.tail=FALSE,df1=df1-df0,df2=n-df1,q=FSTAT)))
  return(ANS)
}
```

Comparison

```
anova(fm0,fm2)
```

```
## Analysis of Variance Table
##
## Model 1: su ~ race + age
## Model 2: su ~ race + race * sex + age
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    397 1019.11
## 2    395  788.36  2    230.75 57.808 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
myANOVA(fm0,fm2)
```

```
##      RES.DF       RSS DF Sum of Sq        F       pValue
## [1,]    397 1019.1072 NA        NA       NA           NA
## [2,]    395  788.3574  2  230.7498 57.80764 9.536695e-23
```

5) Wald's test

Like the F-test, Wald's test can also be used for tests involving 1 or more than 1 df. The test can be used with any null that can be expressed in linear form. The general form of the test is as follows:

- **Ha**: **y=Xb+e** (for this case use your Ha of Q3). Here, **y** is a nx1 vector (the *response*), **X** is an nxp incidence matrix for the pxy vector of effects **b**, and **e** is an nx1 error vector.
- **H0**: **Tb=a**, where **T** is a contrast matrix of dimensions qxp, and **a** is a qx1 vector (often **a=0**).

The covariance matrix of the contrast ($\hat{\mathbf{d}} = \mathbf{T}\hat{\mathbf{b}}$) is $Cov(\hat{\mathbf{d}}) = \mathbf{TCov}(\hat{\mathbf{b}})\mathbf{T}' = \mathbf{S}$, where $Cov(\hat{b})$ is the (co)variance matrix of estimates (Hint: use vcov(fm) to obtain it, here fm is the fitted alternative hypothesis). Note: here $\hat{\mathbf{b}}$ is the OLS estimate of **b** from Ha.

Because of the CLT, in large samples, $\hat{\mathbf{d}} = \mathbf{T}\hat{\mathbf{b}}$ follows a multivariate normal distribution with (co)variance matrix **S**. Therefore, under the null, $(\hat{\mathbf{d}} - \mathbf{a})'\mathbf{S}^{-1}(\hat{\mathbf{d}} - \mathbf{a})$ follows a chi-square distribution with df equal to the rank of **T**.

- Create a function that Implement Wald's test (your function should take a fitted model, representing Ha, and a matrix of contrasts (T). The function should return the test-statistic, test DF, and the p-value.

- Test youf function for the test in 3, compare your p-value with that of the F-test.

```
WTest=function(fm,T){
  DF=min(ncol(T),nrow(T))
  bHat=coef(fm)
  VCOV=vcov(fm)


  if(ncol(T)!=length(bHat)){ stop('The number of columns of T must be equal to the number of parameter
  dHat=T%*%bHat
  V=T%*%VCOV%*%t(T)
  SS=t(dHat)%*%solve(V)%*%dHat
  pvalue=pchisq(q=SS,df=DF,lower.tail=FALSE)
  return(c('df'=DF,'Chisq'=SS,'pvalue'=pvalue))
}

anova(fm0,fm2)
```

```
## Analysis of Variance Table
##
## Model 1: su ~ race + age
## Model 2: su ~ race + race * sex + age
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    397 1019.11
## 2    395  788.36  2    230.75 57.808 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
T=rbind(c(0,0,1,0,0),
        c(0,0,0,0,1))
WTest(fm2,T)
```

```
##           df        Chisq       pvalue
## 2.000000e+00 1.156153e+02 7.842578e-26
```