HW1 Solution

Gustavo

9/29/2021

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
```

Sex and race had highly significant effects on SU, with male and black people having higher SU levels than female and white people, respectively. There is some evidence of an increase in 0.027 units of SU per year of age, but the effect of age is only marginally significant

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.

A non-zero interaction term means that the difference between the average SU levels of male and female varies by race group. It also means that the difference in the average SU levels of white and black people varies between male and female.

- Fit the model with the interaction term, report your results and conclusions.

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

```
##
## Call:
## lm(formula = su ~ race + sex + age + race * sex, 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 main effects of sex and race are highly significant and suggest that, holding everything else constant, male have on average higher SU than female, and that white people have, on average, lower SU levels than black people. The estiamted interaction is positive, suggesting that the difference in SU between whites and blacks is smaller for males, but the p-value is non-significant.Thereofre, we conclude that we have strong evidence of sex and race effects on SU, but we do not have strong evidence in support of an interaction between the two factors

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,

Considering the way the question is formulated, we should allow for the alternative hypothesis to accomodate sex effects that are different for whites and blacks. Thereofre, I choose Ha to be the model with race, sex, age, and the interaction between race and sex. The null hypothesis is the same model without sex on it. Formally:

$Ha: SU_i = \mu + W_i\beta_W + M_i\beta_M + age_i\beta_{age} + (W_i \times M_i)\beta_{WM} + \varepsilon_i$ Versus $H_0: \beta_W = \beta_M = \beta_{WM} = 0$

Above, W_i and M_i are dummy variables for white and male, respectively. Therefore, the product of the two, $W_i \times M_i$ is a dummy variable for white male.

- Test the null using `anova()`, and

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

```
## Analysis of Variance Table
##
## Model 1: su ~ race + age
## Model 2: su ~ race + sex + age + race * sex
##   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
```

- Summarize your findings.

We conclude that sex has a significant 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()`.

```
myANOVA=function(null,alternative){
  n0=length(predict(null))
  na=length(predict(alternative))
  stopifnot(n0==na) # note: we should check that the response is the same, we can only compare models

  p0=length(coef(null))
  pa=length(coef(alternative))

  RSS0=sum(residuals(null)^2)
  RSSA=sum(residuals(alternative)^2)



  ANOVA=matrix(nrow=2,ncol=6,NA)
  rownames(ANOVA)=c('Null','Alternative')
  colnames(ANOVA)=c('Res DF','RSS','DF','SS','F','pvalue')

  ANOVA[1,1]=n0-p0
  ANOVA[2,1]=na-pa

  ANOVA[1,2]=RSS0
  ANOVA[2,2]=RSSA
  ANOVA[2,3]=pa-p0

  df1=(pa-p0)
  df2=(na-pa)
  SS=(RSS0-RSSA)

  ANOVA[2,4]=SS
  FStat=(SS/df1)/(RSSA/df2)
  ANOVA[2,5]=FStat
  ANOVA[2,6]= pf(FStat,df1=df1,df2=df2,lower.tail=FALSE)
  return(ANOVA)
}
```

- Test your function using the H0 and Ha you used in Q3.

```
DATA=read.table('https://raw.githubusercontent.com/gdlc/STAT_COMP/master/DATA/goutData.txt',header=TRU
fm0=lm(su~race+age,data=DATA)
fm2=lm(su~race+sex+age+race*sex,data=DATA)
anova(fm0,fm2)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: su ~ race + age
## Model 2: su ~ race + sex + age + race * sex
##   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       SS        F       pvalue
## Null           397 1019.1072 NA       NA       NA           NA
## Alternative    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 $Cov(\hat{\mathbf{d}}) = \mathbf{T}\mathbf{Cov}(\hat{\mathbf{b}})\mathbf{T}' = \mathbf{S}$, 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.

```
WALD=function(fm,T,a=rep(0,nrow(T)),digits=8){

  Tb=T%*%coef(fm)-a
  COV=T%*%vcov(fm)%*%t(T)
  CHISQ=as.numeric((t(Tb)%*%solve(COV)%*%Tb))
  DF=nrow(T) # or, more precisely, qr(Tb)$rank
  pvalue=pchisq(CHISQ,df=DF,lower.tail=FALSE)
  ANS=round(c('Chisq'=CHISQ,'df'=DF,'pvalue'=pvalue),digits)
  return(ANS)
}
```

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

```
T=rbind(c(0,0,1,0,0),
        c(0,0,0,0,1)
        )
 colnames(T)=names(coef(fm2))
 print(T)
```

```
##      (Intercept) raceW sexM age raceW:sexM
## [1,]           0     0    1   0          0
## [2,]           0     0    0   0          1
```

```
  WALD(fm2,T)
```

```
##    Chisq       df   pvalue
## 115.6153   2.0000   0.0000
```

NOTE: to further test my functions I also tested it in a case where pvalues were not virtually 0

```
 anova(fm1,fm2)
```

```
## Analysis of Variance Table
##
## Model 1: su ~ race + sex + age
## Model 2: su ~ race + sex + age + race * sex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    396 791.06
## 2    395 788.36  1    2.7075 1.3566 0.2448
```

```
 myANOVA(fm1,fm2)
```

```
##             Res DF      RSS DF       SS        F    pvalue
## Null           396 791.0649 NA       NA       NA        NA
## Alternative    395 788.3574  1 2.707493 1.356567 0.2448362
```

```
 T=matrix(c(0,0,0,0,1),nrow=1)
 WALD(fm2,T)
```

```
##     Chisq        df    pvalue
## 1.3565674 1.0000000 0.2441333
```