INF 511 Assignment 4

Jen Diehl (6179236), Sam Watson (6174574), Natasha Wesely (6180693)

2022-10-20

```
longjump_df<- readRDS(file="longjump.RDS")
```

Question 1

a)

```
lmod <- lm(Dist~ RStr + LStr + RFlex + LFlex ,data = longjump_df)
summary(lmod)
```

```
##
## Call:
## lm(formula = Dist ~ RStr + LStr + RFlex + LFlex, data = longjump_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36297 -0.13528 -0.07849  0.09938  0.35893
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.774761   1.032784  13.338 9.55e-07 ***
## RStr         0.005153   0.007645   0.674    0.519
## LStr         0.007697   0.008077   0.953    0.369
## RFlex        0.019404   0.022631   0.857    0.416
## LFlex        0.004614   0.012998   0.355    0.732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2571 on 8 degrees of freedom
## Multiple R-squared:  0.8156, Adjusted R-squared:  0.7235
## F-statistic: 8.848 on 4 and 8 DF,  p-value: 0.004925
```

H_0: There is no collective effect of right leg strength (pounds), left leg strength (pounds), right hamstring flexibility (inches) and left hamstring flexibility (inches) in the high school boys on distance of long jump.

H_a: There is a collective effect of right leg strength (pounds), left leg strength (pounds), right hamstring flexibility (inches) and left hamstring flexibility (inches) in the high school boys on distance of long jump.

The p-value is 0.004925 which is less than alpha, indicating that there is a significant collective effect of all covariates on long jump distance.

b)

We cannot infer a causal effect of covariates on the response because this is observational data that we are assessing statistical trends in, rather than a manipulative experiment where you are explicitly controlling treatments.

c)

We shouldn't be able to infer an association of covariates with the response beyond the sample subjects because that would be interpolating outside of our observed data which is never a good idea.

d)

```
lmod_2 <- lm(Dist ~ 1, data = longjump_df)
anova(lmod, lmod_2)
```

```
## Analysis of Variance Table
##
## Model 1: Dist ~ RStr + LStr + RFlex + LFlex
## Model 2: Dist ~ 1
##   Res.Df     RSS Df Sum of Sq     F   Pr(>F)
## 1      8 0.52871
## 2     12 2.86769 -4    -2.339 8.848 0.004925 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 0.004925 which is less than alpha so we know our more complex model is significantly different from our reduced model. The difference in sums of squares indicates that the reduced model is worse and therefore we should use the full model.

e)

```
#Difference in the number of variables in each model
R = 5- 1
Cmat = cbind(0,diag(R))
d = rep(0,R)
glh.test(reg = lmod, Cmat, d=d)
```

```
##
##   Test of General Linear Hypothesis
## Call:
## glh.test(reg = lmod, cm = Cmat, d = d)
## F = 8.848, df1 = 4, df2 = 8, p-value = 0.004925
```

When using glh.test we get the same p-value and F statistic as when using anova.

f)

H_0: The right and left leg strength have the same effect on (mean) distance jumped.

$\beta_1 = \beta 2$

H_a: The right and left leg strength have different effects on (mean) distance jumped.

β_1 != $\beta 2$

```
legStrMod = lm(Dist~ I(RStr + LStr) + RFlex + LFlex , data = longjump_df)
anova(lmod, legStrMod)
```

```
## Analysis of Variance Table
##
## Model 1: Dist ~ RStr + LStr + RFlex + LFlex
## Model 2: Dist ~ I(RStr + LStr) + RFlex + LFlex
##   Res.Df     RSS Df  Sum of Sq      F Pr(>F)
## 1      8 0.52871
## 2      9 0.53076 -1 -0.0020552 0.0311 0.8644
```

Because there is a high p-value (0.8644), we cannot reject the null hypothesis, meaning we cannot definitively say that the effects are not the same. Therefore, the leg strengths are likely equal. The units for the difference in right and left leg strength are pounds.

g)

```
estimable(lmod, cm=c(0,1,-1,0,0), conf.int = 0.95)
```

```
##                   Estimate Std. Error    t value DF  Pr(>|t|)    Lower.CI
## (0 1 -1 0 0) -0.002543406 0.01442281 -0.1763461  8 0.8644045 -0.03580247
##                 Upper.CI
## (0 1 -1 0 0) 0.03071565
```

We are 95% confident the mean difference of effects (between right and left leg strength) is between -0.03580247 and 0.03071565.

h)

H_0: The right and left leg hamstring flexibility have the same effect on (mean) distance jumped.

$\beta_3 = \beta 4$

H_a: The right and left leg hamstring flexibility have different effects on (mean) distance jumped.

β_3 != $\beta 4$

```
legFlxMod = lm(Dist~ RStr + LStr + I(RFlex + LFlex) , data = longjump_df)
anova(lmod, legFlxMod)
```

```
## Analysis of Variance Table
##
## Model 1: Dist ~ RStr + LStr + RFlex + LFlex
## Model 2: Dist ~ RStr + LStr + I(RFlex + LFlex)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1      8 0.52871
## 2      9 0.54210 -1 -0.013393 0.2027 0.6645
```

```
estimable(lmod, cm=c(0,0,0,1,-1), conf.int = 0.95)
```

```
##                 Estimate Std. Error   t value DF  Pr(>|t|)    Lower.CI
## (0 0 0 1 -1) 0.01479015 0.03285406 0.4501773  8 0.6645315 -0.06097144
##                 Upper.CI
## (0 0 0 1 -1) 0.09055174
```

Similar to the test for leg strength difference, we have observed a high p-value (0.6645) from the test for leg flexibility difference. Because the p-value is high, we fail to reject our null hypothesis. Therefore, the effects of left and right hamstring leg flexibility is likely the same.

We are 95% confident the mean difference of effects (between right and left leg hamstring flexibility) is between -0.06097144 and 0.09055174.

i)

```
# cBeta approach
cmat = matrix(c(0,1,-1,0,0,
                0,0,0,1,-1),
              byrow = T,
              nrow = 2)
d = c(0,0)
gmodels::glh.test(lmod, cmat, d = d)
```

```
##
##   Test of General Linear Hypothesis
## Call:
## gmodels::glh.test(reg = lmod, cm = cmat, d = d)
## F = 0.1175, df1 = 2, df2 = 8, p-value = 0.8907
```

```
# reduced vs full model approach
symmMod = lm(Dist ~ I(RStr + LStr) + I(RFlex + LFlex), data = longjump_df)
anova(lmod, symmMod)
```

```
## Analysis of Variance Table
##
## Model 1: Dist ~ RStr + LStr + RFlex + LFlex
## Model 2: Dist ~ I(RStr + LStr) + I(RFlex + LFlex)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1      8 0.52871
## 2     10 0.54423 -2 -0.015524 0.1175 0.8907
```

Based on the $c\beta$ approach and the full vs reduced model approach above, we cannot reject the null hypothesis (because the p-value is quite large (0.8907)). The reduced model is not significantly different from the full model. Therefore we cannot reject the idea that the legs are symmetrical.

Question 2

```
set.seed(8675309)
nreps <- 5000

fstats <- numeric(nreps)

for (i in 1:nreps){
  lmods <- lm(sample(Dist) ~  RStr + LStr + RFlex + LFlex ,data = longjump_df)
  fstats[i] <- summary(lmods)$fstat[1]
}

mean(fstats> summary(lmod)$fstat[1])
```

```
## [1] 0.0046
```

The p-value from the permutation test (0.0046) is very similar to the p-value from the normal theory test (0.004925). This means that the permutation test supports the original hypothesis test conclusion that there is a collective effect of right leg strength (pounds), left leg strength (pounds), right hamstring flexibility (inches) and left hamstring flexibility (inches) on distance of long jump.

Question 3

```
set.seed(5551212)
nb <- 5000
coefmat <- matrix(NA,nb,5) ## <-- to hold betastar vectors
resids <- residuals(lmod) ## <-- residual vector epshat
preds <- fitted(lmod) ## <-- fitted vector yhat
for(i in 1:nb){
  booty <- preds + sample(resids, rep=TRUE)
  bmod <- update(lmod, booty ~.)
  coefmat[i,] <- coef(bmod)
}

# rename the columns & convert to df for ease
colnames(coefmat) <- c("Intercept",colnames(longjump_df[,2:5]))
coefmat <- data.frame(coefmat)

# get the difference in effects of right & left leg strength
legStrDiff = coefmat[,2] - coefmat[,3]

# 95% empirical CIs for difference in effects of right & left leg strength
quantile(legStrDiff, probs = c(0.025,0.975))
```

```
##        2.5%       97.5%
## -0.02458337  0.01881140
```

For the difference in effects of right and left leg strength, the bootstrap confidence interval is (-0.02458337, 0.01881140). This means we are 95% confident the mean difference in effects is between -0.02458337 and

0.01881140. This bootstrap confidence interval is pretty similar to the normal theory confidence interval (-0.03580247, 0.03071565). The magnitude of difference between the bootstrap CI and normal theory CI is unsurprising given the small sample size.

```
# get the difference in effects of right & left leg hamstring flexibility
legFlxDiff = coefmat[,4] - coefmat[,5]

# 95% empirical CIs for difference in effects of right & left leg hamstring flexibility
quantile(legFlxDiff, probs = c(0.025,0.975))
```

```
##        2.5%       97.5%
## -0.03408454  0.06418265
```

For the difference in effects of right and left leg hamstring flexibility, the bootstrap confidence interval is (-0.03408454, 0.06418265). This means we are 95% confident the mean difference in effects is between -0.03408454 and 0.06418265 . This bootstrap confidence interval is pretty similar to the normal theory confidence interval (-0.06097144, 0.09055174). The magnitude of difference between the bootstrap CI and normal theory CI is unsurprising given the small sample size.

Question 4

a)

```
library(faraway)
data(fat,package="faraway")
lmod_bf <- lm(brozek ~ age + weight + height + neck + chest +
                abdom + hip + thigh + knee + ankle +
                biceps + forearm + wrist, data=fat)
```

```
#E(Y |x0)
x <- model.matrix(lmod_bf)
(x0 <- apply(x,2,mean))
```

```
## (Intercept)         age      weight      height        neck       chest
##     1.00000    44.88492   178.92440    70.14881    37.99206   100.82421
##       abdom         hip       thigh        knee       ankle      biceps
##    92.55595    99.90476    59.40595    38.59048    23.10238    32.27341
##     forearm       wrist
##    28.66389    18.22976
```

```
(y0 <- sum(x0*coef(lmod_bf)))
```

```
## [1] 18.93849
```

```
# 95% prediction interval for E(Y |x0)
(est <- predict(lmod_bf,new=data.frame(t(x0)), interval="confidence", se.fit=TRUE))
```

```
## $fit
##        fit     lwr      upr
## 1 18.93849 18.4436 19.43339
##
## $se.fit
## [1] 0.2512187
##
## $df
## [1] 238
##
## $residual.scale
## [1] 3.987973
```

We are 95% confident the mean expected value for brozek score (E(Y |x0)) is between 18.4436 and 19.43339.

```
(pred<- predict(lmod_bf,new=data.frame(t(x0)), interval="prediction", se.fit=TRUE))
```

```
## $fit
##        fit      lwr     upr
## 1 18.93849 11.06669 26.8103
##
## $se.fit
## [1] 0.2512187
##
## $df
## [1] 238
##
## $residual.scale
## [1] 3.987973
```

For an individual Y|x0, we are 95% confident the mean brozek score is between 11.06669 and 26.8103.

b)

A medical doctor should use a prediction interval because you don't know if that individual patient exhibits normal/average attributes. It would be safer to use the prediction interval.

c)

An exercise science researcher would use the confidence interval to infer about the relationship of percent body fat to these characteristics because they are interested in the general trend in the population which is better represented by the confidence interval around the expected mean value of body fat and characteristics.