Randomization tests - Solutions

The file sablefish.csv contains data from Kimura (1988) on the number of sablefish caught per unit effort (*catch*) in four Alaskan locations for each of the six years between 1978 and 1983.

```
sable <- read.csv("../donnees/sablefish.csv")
head(sable)
```

```
##   year location catch
## 1 1978 Shumagin 0.236
## 2 1978 Chirikof 0.204
## 3 1978   Kodiak 0.241
## 4 1978  Yakutat 0.232
## 5 1979 Shumagin 0.140
## 6 1979 Chirikof 0.202
```

a) Fit a linear model of catch as a function of location only. What is the interpretation of the `locationYakutat` coefficient of this model?

Solution

```
lm_sable <- lm(catch ~ location, sable)
summary(lm_sable)
```

```
##
## Call:
## lm(formula = catch ~ location, data = sable)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17033 -0.09683 -0.04983  0.09471  0.24267
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.33100    0.05483   6.037 6.69e-06 ***
## locationKodiak    0.06733    0.07754   0.868    0.396
## locationShumagin -0.03483    0.07754  -0.449    0.658
## locationYakutat  -0.01167    0.07754  -0.150    0.882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1343 on 20 degrees of freedom
## Multiple R-squared:  0.08762,    Adjusted R-squared:  -0.04924
## F-statistic: 0.6402 on 3 and 20 DF,  p-value: 0.598
```

The coefficient `locationYakutat` gives the mean difference in catch between Yakutat and the reference location (Chirikof).

b) Perform a permutation test to calculate the p-value corresponding to the mean difference in catch between the Kodiak and Chirikof locations. Is this value consistent with the corresponding value in the linear model?

Solution

```
set.seed(8202)

diff_perm <- function() {
   sable_perm <- sable
   sable_perm$location <- sample(sable$location)
   mean(sable_perm$catch[sable_perm$location == "Kodiak"]) -
       mean(sable_perm$catch[sable_perm$location == "Chirikof"])
}

nperm <- 9999

diff_null <- replicate(nperm, diff_perm())

diff_obs <- mean(sable$catch[sable$location == "Kodiak"]) -
            mean(sable$catch[sable$location == "Chirikof"])

(sum(abs(diff_null) >= abs(diff_obs)) + 1) / (nperm + 1)
```

```
## [1] 0.3823
```

Yes, the value is close to the p-value for the `locationKodiak` coefficient of the model in a).

c) Using the *permuco* package, determine the p-value for the same difference, for a model including the additive effects of year and location. *Note*: We consider the year as a categorical variable here, so it must be converted to a factor. Does the p-value differ between the permutation test and the parametric model?

Solution

```
library(permuco)
sable$year <- as.factor(sable$year)
lmperm(catch ~ year + location, sable)
```

```
## Table of marginal t-test of the betas
## Permutation test using freedman_lane to handle nuisance variables and 5000 permutations.
##                  Estimate Std. Error t value parametric Pr(>|t|)
## (Intercept)       0.22304    0.03849  5.7944           3.537e-05
## year1979         -0.01875    0.04445 -0.4218           6.791e-01
## year1980          0.06300    0.04445  1.4174           1.768e-01
## year1981          0.10725    0.04445  2.4130           2.908e-02
## year1982          0.30450    0.04445  6.8508           5.499e-06
## year1983          0.19175    0.04445  4.3141           6.142e-04
## locationKodiak    0.06733    0.03629  1.8554           8.330e-02
## locationShumagin -0.03483    0.03629 -0.9598           3.524e-01
## locationYakutat  -0.01167    0.03629 -0.3215           7.523e-01
##                  permutation Pr(<t) permutation Pr(>t) permutation Pr(>|t|)
## (Intercept)
## year1979                     0.3286             0.6716               0.6778
## year1980                     0.9044             0.0958               0.1828
## year1981                     0.9890             0.0112               0.0236
## year1982                     1.0000             0.0002               0.0002
## year1983                     0.9996             0.0006               0.0006
## locationKodiak               0.9642             0.0360               0.0794
## locationShumagin             0.1824             0.8178               0.3476
## locationYakutat              0.3836             0.6166               0.7520
```

The p-value is about 0.08 for both the parametric model and the permutation test.