HW2.1.5 // Abalone Regression

Code

```
set.seed(2020)

ABALONE_FEATURES = c(
  'length',
  'diameter',
  'height',
  'whole_weight',
  'shucked_weight',
  'viscera_weight',
  'shell_weight',
  'is_m',
  'is_f',
  'is_i',
  "x1",
  "x2"
)

data = read.csv("~/Stanford/STATS305A/abalone_data.csv") %>%
  mutate(is_m = ifelse(sex == "M", 1, 0),
         is_f = ifelse(sex == "F", 1, 0),
         is_i = ifelse(sex == "I", 1, 0))

# add noise per rho
rho = .1


do_single_experiment <- function(data, rho){
  data_noise = cbind(
    data,
    data.frame(
      rmvnorm(nrow(data), c(0, 0), matrix(c(1, rho, rho, 1), ncol=2))
    ) %>%
      `colnames<-`(c("x1", "x2"))
  )

  # fit model
  fit = lm(
    formula = as.formula(paste0("rings ~ ", paste0(ABALONE_FEATURES, collapse = " + "))),
    data = data_noise
  ) %>%
    summary()
```

```r
  # extract pvals
  pval_x1 = fit$coefficients[, 4] %>% .[length(.) - 1] %>% as.numeric()
  pval_x2 = fit$coefficients[, 4] %>% .[length(.)] %>% as.numeric()

  c(pval_x1, pval_x2)
}


do_experiments <- function(data, rho, n_exp=1000, alpha=.05){
  lapply(1:n_exp, function(i) do_single_experiment(data=data, rho=rho)) %>%
    do.call("rbind",. ) %>%
    data.frame() %>%
    `colnames<-`(c("p1", "p2"))%>%
    mutate(reject_1 = ifelse(p1 < alpha, 1, 0),
           reject_2 = ifelse(p2 < alpha, 1, 0),
           reject_dual = reject_1 * reject_2) -> sim_result
  sim_result[, c("reject_1", "reject_2", "reject_dual")] %>%
    colSums()
}


RHO_VEC = c(-.9, -.8, -.4, 0, .4, .8, .9)

lapply(RHO_VEC, function(rho) do_experiments(data, rho=rho)) %>%
  do.call("rbind", .) %>%
  data.frame() %>%
  `colnames<-`(c("reject_1", "reject_2", "reject_dual")) -> full_result
full_result$rho = RHO_VEC



full_result
```

```
##   reject_1 reject_2 reject_dual  rho
## 1       44       43          28 -0.9
## 2       45       48          19 -0.8
## 3       53       55           6 -0.4
## 4       50       43           1  0.0
## 5       59       54           6  0.4
## 6       43       38          17  0.8
## 7       54       46          32  0.9
```

Indeed, we see the expected clumpiness that stems from the correlation added to X1 and X2 in the design matrix . While marginally the rejections look even (roughly 40-60) per 1000, we see that dual rejections increase the farther rho gets from 0 – closer to rho = 0, there are very few dual rejections, while at rho s.t. |rho| = .9, there are around 30.