module\_6\_assignment\_code

library(MPV)  
  
### Problem 7.6  
orig\_df <- MPV::p7.6  
n <- nrow(orig\_df)  
ones <- rep(1, n)  
y <- orig\_df[,1]  
x1 <- orig\_df[,2]  
x2 <- orig\_df[,3]  
  
# part (a) Fit a second-order polynomial.  
df\_a <- cbind(ones, x1, x2, x1^2, x2^2, x1\*x2)  
  
beta\_hat\_calc <- function(X, y) {  
 beta\_hat <- solve(t(X) %\*% X) %\*% t(X) %\*% y  
 return(beta\_hat)  
}  
H\_calc <- function(X) {  
 H <- X %\*% solve(t(X) %\*% X) %\*% t(X)  
 return(H)  
}  
y\_hat\_calc <- function(H, y) {  
 y\_hat <- H %\*% y  
 return(y\_hat)  
}  
e\_calc <- function(y, y\_hat) {  
 e <- y - y\_hat  
 return(e)  
}  
  
beta\_hat\_a <- beta\_hat\_calc(X = df\_a, y = y)  
H\_a <- H\_calc(X = df\_a)  
y\_hat\_a <- y\_hat\_calc(H = H\_a, y = y)  
e\_a <- e\_calc(y = y, y\_hat = y\_hat\_a)  
t(e\_a) %\*% e\_a # 2.302162  
  
# part (b) Test for significance of regression.  
SS\_Res\_calc <- function(y, beta\_hat, X) {  
 SS\_Res <- (t(y) %\*% y) - (t(beta\_hat) %\*% t(X) %\*% y)  
 return(SS\_Res)  
}  
SS\_T\_calc <- function(y) {  
 n <- length(y)  
 SS\_T <- (t(y) %\*% y) - ((sum(y)^2) / n)  
 return(SS\_T)  
}  
SS\_R\_calc <- function(beta\_hat, X, y) {  
 n <- length(y)  
 SS\_R <- (t(beta\_hat) %\*% t(X) %\*% y) - ((sum(y)^2) / n)  
 return(SS\_R)  
}  
  
SS\_Res\_a <- SS\_Res\_calc(beta\_hat = beta\_hat\_a, X = df\_a, y = y)  
SS\_T\_a <- SS\_T\_calc(y = y)  
SS\_R\_a <- SS\_R\_calc(beta\_hat = beta\_hat\_a, X = df\_a, y = y)  
  
SS\_Res\_a == SS\_T\_a - SS\_R\_a  
SS\_Res\_a; SS\_T\_a; SS\_R\_a # 2.302142, 342.1899, 339.8878  
k <- 5; n - k - 1; n - 1  
MS\_R\_a <- SS\_R\_a / k # 67.97755  
MS\_Res\_a <- SS\_Res\_a / (n - k - 1) # 0.3836904  
F\_a <- MS\_R\_a / MS\_Res\_a # 177.1677  
  
alpha <- 0.1  
qf(p = (1 - alpha), df1 = k, df2 = (n - k - 1))  
pf(q = F\_a, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)  
  
# Double-check with lm / anova  
m1 <- lm(y ~ x1\*x2 + I(x1^2) + I(x2^2))  
m2 <- lm(y~1)  
anova(m1, m2)  
  
# part (d) Does the interaction term contribute significantly to the model?  
C\_mat\_calc <- function(X) {  
 C\_mat <- solve(t(X) %\*% X)  
 return(C\_mat)  
}  
  
sigma\_hat\_squared <- MS\_Res\_a  
C\_mat\_a <- C\_mat\_calc(X = df\_a)  
t\_a <- (beta\_hat\_a[6] / sqrt(sigma\_hat\_squared \* C\_mat\_a[6,6]))  
  
alpha <- 0.05  
qt(p = (1 - alpha / 2), df = (n - k - 1))  
  
# Reference: https://stats.stackexchange.com/questions/45153/manually-calculating-p-value-from-t-value-in-t-test  
2 \* pt(q = abs(t\_a), df = (n - k - 1), lower.tail = FALSE)  
  
# Double-check with summary  
summary(m1)  
  
# part (e) Do the second-order terms contribute significantly to the model?  
df\_a1 <- df\_a[,c(1,2,3)]  
beta\_hat\_red <- beta\_hat\_calc(X = df\_a1, y = y)  
SS\_R\_b1\_a <- SS\_R\_calc(beta\_hat = beta\_hat\_red, X = df\_a1, y = y)  
SS\_R\_b2\_given\_b1\_a <- SS\_R\_a - SS\_R\_b1\_a  
  
r <- 3  
F\_a1 <- (SS\_R\_b2\_given\_b1\_a / r) / MS\_Res\_a # 5.057906  
alpha <- 0.05  
qf(p = (1 - alpha), df1 = r, df2 = (n - k - 1))  
pf(q = F\_a1, df1 = r, df2 = (n - k - 1), lower.tail = FALSE)  
  
# Double-check  
m2 <- lm(y ~ x1 + x2)  
anova(m1, m2)  
  
### Problem 8.11  
orig\_df <- MPV::p8.11  
n <- nrow(orig\_df)  
y <- orig\_df[,1]  
Xs <- orig\_df[,2]  
ones <- rep(1, n)  
  
indicator\_ones <- rep(1, 5)  
indicator\_zeros <- rep(0, 5)  
x1 <- c(indicator\_ones, rep(indicator\_zeros, 4))  
x2 <- c(rep(indicator\_zeros, 1), indicator\_ones, rep(indicator\_zeros, 3))  
x3 <- c(rep(indicator\_zeros, 2), indicator\_ones, rep(indicator\_zeros, 2))  
x4 <- c(rep(indicator\_zeros, 3), indicator\_ones, rep(indicator\_zeros, 1))  
  
X <- cbind(ones, x1, x2, x3, x4)  
  
# part (b) Find the least-squares estimates of the model parameters.  
beta\_hat <- beta\_hat\_calc(X = X, y = y)  
H <- H\_calc(X = X)  
y\_hat <- y\_hat\_calc(H = H, y = y)  
e <- e\_calc(y = y, y\_hat = y\_hat)  
t(e) %\*% e # 161.2  
  
# Double-check  
m1 <- lm(y~X-1)  
  
# part (c) Find a point estimate of the difference in mean strength between 15% and 25% cotton.  
beta\_hat[2] - beta\_hat[4] # -7.8  
  
# part (d) Test the hypothesis that the mean tensile strength is the same for all five cotton percentages.  
k <- 4  
SS\_R\_full <- SS\_R\_calc(beta\_hat = beta\_hat, X = X, y = y)  
MS\_R\_full <- SS\_R\_full / k # 118.94  
SS\_Res <- SS\_Res\_calc(beta\_hat = beta\_hat, X = X, y = y)  
MS\_Res <- SS\_Res / (n - k - 1) # 8.06  
F\_0 <- MS\_R\_full / MS\_Res # 14.75682  
pf(q = F\_0, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)  
  
# Double-check  
X\_red <- X[,1]  
m2 <- lm(y~X\_red-1)  
anova(m1, m2)  
  
### Problem 8.16  
Location <- seq(1, 17)  
INHIBIT <- c(0.00, 1.00, 6.00, 7.00, 7.00, 7.00, 9.00, 9.50, 10.00, 11.00, 12.50, 14.00, 20.00, 21.00, 25.00, 39.00, 59.00)  
UVB <- c(0.00, 0.00, 0.01, 0.01, 0.02, 0.03, 0.04, 0.01, 0.00, 0.03, 0.03, 0.01, 0.03, 0.04, 0.02, 0.03, 0.03)  
SURFACE <- c('Deep', 'Deep', 'Deep', 'Surface', 'Surface', 'Surface', 'Surface', 'Deep', 'Deep', 'Surface', 'Surface', 'Deep', 'Deep', 'Surface', 'Deep', 'Deep', 'Deep')  
orig\_df <- data.frame(Location, INHIBIT, UVB, SURFACE)  
  
set.seed(1); chosen\_rows <- sample(Location, 12)  
chosen\_rows <- sort(chosen\_rows)  
n <- length(chosen\_rows)  
ones <- rep(1, n)  
df <- orig\_df[chosen\_rows,]  
y <- df$INHIBIT  
x1 <- df$UVB  
  
# Reference: https://stackoverflow.com/questions/40780088/r-code-categorical-variable-to-1-and-0  
df$SURFACE <- as.factor(df$SURFACE)  
df$is\_Deep <- as.numeric(df$SURFACE)  
df[df$is\_Deep == 2,]$is\_Deep <- 0  
x2 <- df$is\_Deep  
X <- cbind(ones, x1, x2, x1\*x2)  
  
# Test for significance of regression  
k <- 3  
beta\_hat <- beta\_hat\_calc(X = X, y = y)  
SS\_R\_full <- SS\_R\_calc(beta\_hat = beta\_hat, X = X, y = y)  
MS\_R\_full <- SS\_R\_full / k  
SS\_Res <- SS\_Res\_calc(beta\_hat = beta\_hat, X = X, y = y)  
MS\_Res <- SS\_Res / (n - k - 1)  
F\_0 <- MS\_R\_full / MS\_Res # 9.119374  
pf(q = F\_0, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)  
  
# Double-check  
m1 <- lm(y ~ x1\*x2); m2 <- lm(y ~ 1)  
anova(m1, m2)  
  
# Test for interaction term significance  
sigma\_hat\_squared <- MS\_Res  
C\_mat\_a <- C\_mat\_calc(X = X)  
t\_0 <- (beta\_hat[4] / sqrt(sigma\_hat\_squared \* C\_mat\_a[4,4]))  
  
alpha <- 0.05  
qt(p = (1 - alpha / 2), df = (n - k - 1))  
  
2 \* pt(q = abs(t\_0), df = (n - k - 1), lower.tail = FALSE)  
  
# Double-check with summary  
summary(m1)  
  
# Test b1 and b2  
t\_0\_b1 <- (beta\_hat[2] / sqrt(sigma\_hat\_squared \* C\_mat\_a[2,2]))  
t\_0\_b2 <- (beta\_hat[3] / sqrt(sigma\_hat\_squared \* C\_mat\_a[3,3]))  
  
2 \* pt(q = abs(t\_0\_b1), df = (n - k - 1), lower.tail = FALSE)  
2 \* pt(q = abs(t\_0\_b2), df = (n - k - 1), lower.tail = FALSE)