module\_8\_assignment\_code

library(MPV)  
  
### Problem 1  
n <- 20  
set.seed(1); chosen\_rows <- sort(sample(seq(1,27), n))  
df <- MPV::table.b5  
df <- df[chosen\_rows,c(1, 2, 7)]  
ones <- rep(1, n)  
y <- df[,1]  
X <- cbind(ones, df[,c(2,3)])  
  
beta\_hat\_calc <- function(X, y) {  
 X <- as.matrix(X)  
 beta\_hat <- solve(t(X) %\*% X) %\*% t(X) %\*% y  
 return(beta\_hat)  
}  
H\_calc <- function(X) {  
 X <- as.matrix(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 <- beta\_hat\_calc(X=X, y=y)  
H <- H\_calc(X)  
y\_hat <- y\_hat\_calc(H=H, y=y)  
e <- e\_calc(y=y, y\_hat=y\_hat)  
  
#### part (a)  
norm\_prob\_plot <- function(residual\_var, x\_label,  
 main\_title = 'Normal Probability Plot',  
 y\_label = 'Probability', n\_size=n) {  
 ones <- rep(1, n)  
 sorted\_residuals <- sort(residual\_var)  
 cumulative\_probability <- (1:n\_size - 0.5) / n\_size  
 plot(sorted\_residuals, cumulative\_probability, main = main\_title,  
 xlab = x\_label,  
 ylab = y\_label)  
 X\_temp <- cbind(ones, sorted\_residuals)  
 beta\_hat\_temp <- beta\_hat\_calc(X=X\_temp,y=cumulative\_probability)  
 abline(beta\_hat\_temp)  
}  
norm\_prob\_plot(residual\_var = e, x\_label = 'Sorted Residuals')  
  
order(e, decreasing = FALSE)  
e[order(e, decreasing = FALSE)]  
  
### part (b)  
res\_vs\_fitted\_plot <- function(residual\_var,  
 main\_title,  
 y\_label,  
 x\_label = 'Predicted Response',  
 pred\_response = y\_hat) {  
 plot(pred\_response, residual\_var, main = main\_title,  
 xlab = x\_label,  
 ylab = y\_label,  
 ylim = c(min(residual\_var)-sd(residual\_var),  
 max(residual\_var)+sd(residual\_var)))  
}  
res\_vs\_fitted\_plot(residual\_var = e,  
 main\_title = 'Residuals vs. Predicted Response',  
 y\_label = 'Residuals')  
  
### part (c)  
# studentized residuals  
SS\_Res\_calc <- function(y, beta\_hat, X) {  
 SS\_Res <- (t(y) %\*% y) - (t(beta\_hat) %\*% t(X) %\*% y)  
 return(SS\_Res)  
}  
SS\_Res <- SS\_Res\_calc(y = y, beta\_hat = beta\_hat, X = X)  
p <- ncol(X)  
MS\_Res <- SS\_Res / (n - p)  
H\_diag <- diag(H)  
studentized\_residuals <- sapply(1:n, function(x) {  
 e[x] / sqrt(MS\_Res \* (1 - H\_diag[x]))  
})  
  
# res vs. fitted, norm prob res, res vs. x1, res vs. x6  
res\_vs\_regressor <- function(residual\_var,  
 main\_title1, main\_title2,  
 ylabel,  
 X\_df=X) {  
 x1 <- X\_df[,2]; x6 <- X\_df[,3]  
 plot(x1, residual\_var,  
 main = main\_title1,  
 ylab = ylabel,  
 xlab = 'Space time, min.')  
 abline(h = 0)  
  
 plot(x6, residual\_var,  
 main = main\_title2,  
 ylab = ylabel,  
 xlab = 'Solvent total')  
 abline(h = 0)  
}  
  
par(mfrow = c(2,2))  
res\_vs\_fitted\_plot(residual\_var = studentized\_residuals,  
 main\_title = 'Studentized Residuals vs. Predicted Response',  
 y\_label = 'Studentized Residuals')  
norm\_prob\_plot(residual\_var = studentized\_residuals, x\_label = 'Studentized Residuals')  
res\_vs\_regressor(residual\_var = studentized\_residuals,  
 main\_title1 = 'Studentized Residuals vs. Space time, min',  
 main\_title2 = 'Studentized Residuals vs. Solvent total',  
 ylabel = 'Studentized Residuals')  
  
# Analyze the outliers  
head(order(x6, decreasing = TRUE), 5)  
studentized\_residuals[12] # bottom right  
studentized\_residuals[1]  
studentized\_residuals[18]  
studentized\_residuals[8] # top right  
studentized\_residuals[11]  
  
# R-student  
S\_squared <- sapply(1:n, function(x) {  
 ((n - p) \* MS\_Res - ((e[x]^2) / (1 - H\_diag[x]))) / (n - p - 1)  
})  
  
R\_student\_res <- sapply(1:n, function(x) {  
 e[x] / sqrt(S\_squared[x] \* (1 - H\_diag[x]))  
})  
  
par(mfrow = c(2,2))  
res\_vs\_fitted\_plot(residual\_var = R\_student\_res,  
 main\_title = 'R-student vs. Predicted Response',  
 y\_label = 'R-student')  
norm\_prob\_plot(residual\_var = R\_student\_res, x\_label = 'R-student')  
res\_vs\_regressor(residual\_var = R\_student\_res,  
 main\_title1 = 'R-student vs. Space time, min',  
 main\_title2 = 'R-student vs. Solvent total',  
 ylabel = 'R-student')  
  
### part (d)  
# standardized residuals  
standardized\_res <- sapply(1:n, function(x) e[x] / sqrt(MS\_Res))  
  
par(mfrow = c(2,2))  
res\_vs\_fitted\_plot(residual\_var = standardized\_res,  
 main\_title = 'Standardized Residuals vs. Predicted Response',  
 y\_label = 'Standardized Residuals')  
norm\_prob\_plot(residual\_var = standardized\_res, x\_label = 'Standardized Residuals')  
res\_vs\_regressor(residual\_var = standardized\_res,  
 main\_title1 = 'Standardized Residuals vs. Space time, min',  
 main\_title2 = 'Standardized Residuals vs. Solvent total',  
 ylabel = 'Standardized Residuals')  
  
  
# PRESS residuals  
PRESS\_res <- sapply(1:n, function(x) e[x] / (1 - H\_diag[x]))  
  
par(mfrow = c(2,2))  
res\_vs\_fitted\_plot(residual\_var = PRESS\_res,  
 main\_title = 'PRESS Residuals vs. Predicted Response',  
 y\_label = 'PRESS Residuals')  
norm\_prob\_plot(residual\_var = PRESS\_res, x\_label = 'PRESS Residuals')  
res\_vs\_regressor(residual\_var = PRESS\_res,  
 main\_title1 = 'PRESS Residuals vs. Space time, min',  
 main\_title2 = 'PRESS Residuals vs. Solvent total',  
 ylabel = 'PRESS Residuals')  
  
  
### Problem 2  
df <- MPV::table.b4  
set.seed(2); chosen\_rows <- sort(sample(seq(1,24), 15))  
df <- df[chosen\_rows,c(1, 5, 8, 10)]  
n <- nrow(df)  
ones <- rep(1, n)  
y <- df[,1]  
X <- cbind(ones, df[,c(2:4)])  
  
beta\_hat <- beta\_hat\_calc(X=X, y=y)  
H <- H\_calc(X)  
y\_hat <- y\_hat\_calc(H=H, y=y)  
e <- e\_calc(y=y, y\_hat=y\_hat)  
### part (a)  
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 <- SS\_Res\_calc(beta\_hat = beta\_hat, X = X, y = y)  
SS\_T <- SS\_T\_calc(y = y)  
SS\_R <- SS\_R\_calc(beta\_hat = beta\_hat, X = X, y = y)  
  
SS\_Res == SS\_T - SS\_R  
SS\_Res; SS\_T; SS\_R  
k <- 3; p <- k + 1  
MS\_R <- SS\_R / k  
MS\_Res <- SS\_Res / (n - k - 1)  
F\_0 <- MS\_R / MS\_Res  
  
alpha <- 0.05  
qf(p = (1 - alpha), df1 = k, df2 = (n - k - 1))  
pf(q = F\_0, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)  
  
r\_squared\_calc <- function(SS\_R, SS\_T) {  
 r\_squared <- SS\_R / SS\_T  
 return(r\_squared)  
}  
adj\_r\_squared\_calc <- function(SS\_Res, SS\_T, n, p) {  
 adj\_r\_squared <- 1 - ((SS\_Res / (n - p)) / (SS\_T / (n - 1)))  
 return(adj\_r\_squared)  
}  
  
r\_squared <- r\_squared\_calc(SS\_R = SS\_R, SS\_T = SS\_T)  
adj\_r\_squared <- adj\_r\_squared\_calc(  
 SS\_Res = SS\_Res, SS\_T = SS\_T, n = n, p = p)  
  
# ordinary residuals  
par(mfrow = c(2,2))  
norm\_prob\_plot(residual\_var = e, x\_label = 'Sorted Residuals')  
res\_vs\_fitted\_plot(residual\_var = e,  
 main\_title = 'Residuals vs. Predicted Response',  
 y\_label = 'Residuals')  
  
order(e, decreasing = TRUE)  
  
# studentized residuals  
H\_diag <- diag(H)  
studentized\_residuals <- sapply(1:n, function(x) {  
 e[x] / sqrt(MS\_Res \* (1 - H\_diag[x]))  
})  
norm\_prob\_plot(residual\_var = studentized\_residuals, x\_label = 'Studentized Residuals')  
res\_vs\_fitted\_plot(residual\_var = studentized\_residuals,  
 main\_title = 'Studentized Residuals vs. Predicted Response',  
 y\_label = 'Studentized Residuals')  
  
# R-student residuals  
par(mfrow = c(2,2))  
S\_squared <- sapply(1:n, function(x) {  
 ((n - p) \* MS\_Res - ((e[x]^2) / (1 - H\_diag[x]))) / (n - p - 1)  
})  
  
R\_student\_res <- sapply(1:n, function(x) {  
 e[x] / sqrt(S\_squared[x] \* (1 - H\_diag[x]))  
})  
  
norm\_prob\_plot(residual\_var = R\_student\_res, x\_label = 'R-student')  
res\_vs\_fitted\_plot(residual\_var = R\_student\_res,  
 main\_title = 'R-student vs. Predicted Response',  
 y\_label = 'R-student')  
  
# standardized residuals  
standardized\_res <- sapply(1:n, function(x) e[x] / sqrt(MS\_Res))  
norm\_prob\_plot(residual\_var = standardized\_res, x\_label = 'Standardized Residuals')  
res\_vs\_fitted\_plot(residual\_var = standardized\_res,  
 main\_title = 'Standardized Residuals vs. Predicted Response',  
 y\_label = 'Standardized Residuals')  
  
# PRESS residuals  
par(mfrow = c(2,2))  
PRESS\_res <- sapply(1:n, function(x) e[x] / (1 - H\_diag[x]))  
norm\_prob\_plot(residual\_var = PRESS\_res, x\_label = 'PRESS Residuals')  
res\_vs\_fitted\_plot(residual\_var = PRESS\_res,  
 main\_title = 'PRESS Residuals vs. Predicted Response',  
 y\_label = 'PRESS Residuals')  
  
x4 <- X[,2]; x7 <- X[,3]; x9 <- X[,4]  
par(mfrow = c(2,2))  
plot(x4, e,  
 main = 'Residuals vs. Living Space (sq ft x 1000)',  
 ylab = 'Residuals',  
 xlab = 'Living Space (sq ft x 1000)')  
abline(h = 0)  
  
plot(x7, e,  
 main = 'Residuals vs. Number of Bedrooms',  
 ylab = 'Residuals',  
 xlab = 'Number of Bedrooms')  
abline(h = 0)  
  
plot(x9, e,  
 main = 'Residuals vs. Number of Fireplaces',  
 ylab = 'Residuals',  
 xlab = 'Number of Fireplaces')  
abline(h = 0)  
  
  
### part (b)  
data.frame(table(y))  
  
### Problem 3  
### part (a)  
df <- MPV::p5.5  
set.seed(3); chosen\_rows <- sort(sample(seq(1,14), 7))  
df <- df[chosen\_rows,]  
y <- df[,1]  
n <- nrow(df)  
ones <- rep(1, n)  
X <- cbind(ones, df[,2])  
  
beta\_hat <- beta\_hat\_calc(X=X, y=y)  
H <- H\_calc(X)  
y\_hat <- y\_hat\_calc(H=H, y=y)  
e <- e\_calc(y=y, y\_hat=y\_hat)  
  
SS\_Res <- SS\_Res\_calc(beta\_hat = beta\_hat, X = X, y = y)  
SS\_T <- SS\_T\_calc(y = y)  
SS\_R <- SS\_R\_calc(beta\_hat = beta\_hat, X = X, y = y)  
  
SS\_Res == SS\_T - SS\_R  
SS\_Res; SS\_T; SS\_R  
k <- 1; p <- k + 1  
MS\_R <- SS\_R / k  
MS\_Res <- SS\_Res / (n - k - 1)  
F\_0 <- MS\_R / MS\_Res  
  
alpha <- 0.01  
qf(p = (1 - alpha), df1 = k, df2 = (n - k - 1))  
pf(q = F\_0, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)  
  
r\_squared <- r\_squared\_calc(SS\_R = SS\_R, SS\_T = SS\_T)  
  
# ordinary residuals  
par(mfrow = c(2,2))  
norm\_prob\_plot(residual\_var = e, x\_label = 'Sorted Residuals')  
res\_vs\_fitted\_plot(residual\_var = e,  
 main\_title = 'Residuals vs. Predicted Response',  
 y\_label = 'Residuals')  
  
order(e, decreasing = TRUE)  
  
# studentized residuals  
H\_diag <- diag(H)  
studentized\_residuals <- sapply(1:n, function(x) {  
 e[x] / sqrt(MS\_Res \* (1 - H\_diag[x]))  
})  
norm\_prob\_plot(residual\_var = studentized\_residuals, x\_label = 'Studentized Residuals')  
res\_vs\_fitted\_plot(residual\_var = studentized\_residuals,  
 main\_title = 'Studentized Residuals vs. Predicted Response',  
 y\_label = 'Studentized Residuals')  
  
# R-student residuals  
par(mfrow = c(2,2))  
S\_squared <- sapply(1:n, function(x) {  
 ((n - p) \* MS\_Res - ((e[x]^2) / (1 - H\_diag[x]))) / (n - p - 1)  
})  
  
R\_student\_res <- sapply(1:n, function(x) {  
 e[x] / sqrt(S\_squared[x] \* (1 - H\_diag[x]))  
})  
  
norm\_prob\_plot(residual\_var = R\_student\_res, x\_label = 'R-student')  
res\_vs\_fitted\_plot(residual\_var = R\_student\_res,  
 main\_title = 'R-student vs. Predicted Response',  
 y\_label = 'R-student')  
  
# standardized residuals  
standardized\_res <- sapply(1:n, function(x) e[x] / sqrt(MS\_Res))  
norm\_prob\_plot(residual\_var = standardized\_res, x\_label = 'Standardized Residuals')  
res\_vs\_fitted\_plot(residual\_var = standardized\_res,  
 main\_title = 'Standardized Residuals vs. Predicted Response',  
 y\_label = 'Standardized Residuals')  
  
# PRESS residuals  
par(mfrow = c(2,2))  
PRESS\_res <- sapply(1:n, function(x) e[x] / (1 - H\_diag[x]))  
norm\_prob\_plot(residual\_var = PRESS\_res, x\_label = 'PRESS Residuals')  
res\_vs\_fitted\_plot(residual\_var = PRESS\_res,  
 main\_title = 'PRESS Residuals vs. Predicted Response',  
 y\_label = 'PRESS Residuals')  
  
par(mfrow = c(2,2))  
plot(X[,2], e,  
 main = 'Residuals vs. Weeks',  
 ylab = 'Residuals',  
 xlab = 'Weeks')  
abline(h = 0)  
  
plot(X[,2], studentized\_residuals,  
 main = 'Studentized Residuals vs. Weeks',  
 ylab = 'Studentized Residuals',  
 xlab = 'Weeks')  
abline(h = 0)  
  
plot(X[,2], R\_student\_res,  
 main = 'R-Student vs. Weeks',  
 ylab = 'R-Student',  
 xlab = 'Weeks')  
abline(h = 0)  
  
plot(X[,2], standardized\_res,  
 main = 'Standardized Residuals vs. Weeks',  
 ylab = 'Standardized Residuals',  
 xlab = 'Weeks')  
abline(h = 0)  
  
par(mfrow = c(1,1))  
plot(X[,2], PRESS\_res,  
 main = 'PRESS Residuals vs. Weeks',  
 ylab = 'PRESS Residuals',  
 xlab = 'Weeks')  
abline(h = 0)  
  
### part (b)  
par(mfrow = c(2,2))  
plot(df$weeks, df$defects,  
 main = 'Defects per 10,000 vs. Weeks',  
 xlab = 'Weeks', ylab = 'Defects per 10,000')  
abline(beta\_hat)  
  
log\_y <- log(df$defects)  
log\_x <- log(df$weeks)  
plot(log\_x, log\_y,  
 main = 'log Defects per 10,000 vs. log Weeks',  
 xlab = 'log Weeks', ylab = 'log Defects per 10,000')  
X\_log <- cbind(ones, log\_x)  
beta\_hat\_log <- beta\_hat\_calc(X=X\_log, y=log\_y)  
H\_log <- H\_calc(X\_log)  
y\_hat\_log <- y\_hat\_calc(H=H\_log, y=log\_y)  
e\_log <- e\_calc(y=log\_y, y\_hat=y\_hat\_log)  
abline(beta\_hat\_log)  
  
SS\_Res <- SS\_Res\_calc(beta\_hat = beta\_hat\_log, X = X\_log, y = log\_y)  
SS\_T <- SS\_T\_calc(y = log\_y)  
SS\_R <- SS\_R\_calc(beta\_hat = beta\_hat\_log, X = X\_log, y = log\_y)  
  
SS\_Res == SS\_T - SS\_R  
SS\_Res; SS\_T; SS\_R  
k <- 1; p <- k + 1  
MS\_R <- SS\_R / k  
MS\_Res <- SS\_Res / (n - k - 1)  
F\_0 <- MS\_R / MS\_Res  
  
alpha <- 0.01  
qf(p = (1 - alpha), df1 = k, df2 = (n - k - 1))  
pf(q = F\_0, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)  
  
r\_squared <- r\_squared\_calc(SS\_R = SS\_R, SS\_T = SS\_T)  
  
# ordinary residuals  
par(mfrow = c(2,2))  
e <- e\_log  
norm\_prob\_plot(residual\_var = e, x\_label = 'Sorted Residuals')  
res\_vs\_fitted\_plot(residual\_var = e,  
 main\_title = 'Residuals vs. Predicted Response',  
 y\_label = 'Residuals')  
  
order(e, decreasing = TRUE)  
  
# studentized residuals  
H\_diag <- diag(H)  
studentized\_residuals <- sapply(1:n, function(x) {  
 e[x] / sqrt(MS\_Res \* (1 - H\_diag[x]))  
})  
norm\_prob\_plot(residual\_var = studentized\_residuals, x\_label = 'Studentized Residuals')  
res\_vs\_fitted\_plot(residual\_var = studentized\_residuals,  
 main\_title = 'Studentized Residuals vs. Predicted Response',  
 y\_label = 'Studentized Residuals')  
  
# R-student residuals  
par(mfrow = c(2,2))  
S\_squared <- sapply(1:n, function(x) {  
 ((n - p) \* MS\_Res - ((e[x]^2) / (1 - H\_diag[x]))) / (n - p - 1)  
})  
  
R\_student\_res <- sapply(1:n, function(x) {  
 e[x] / sqrt(S\_squared[x] \* (1 - H\_diag[x]))  
})  
  
norm\_prob\_plot(residual\_var = R\_student\_res, x\_label = 'R-student')  
res\_vs\_fitted\_plot(residual\_var = R\_student\_res,  
 main\_title = 'R-student vs. Predicted Response',  
 y\_label = 'R-student')  
  
# standardized residuals  
standardized\_res <- sapply(1:n, function(x) e[x] / sqrt(MS\_Res))  
norm\_prob\_plot(residual\_var = standardized\_res, x\_label = 'Standardized Residuals')  
res\_vs\_fitted\_plot(residual\_var = standardized\_res,  
 main\_title = 'Standardized Residuals vs. Predicted Response',  
 y\_label = 'Standardized Residuals')  
  
# PRESS residuals  
par(mfrow = c(2,2))  
PRESS\_res <- sapply(1:n, function(x) e[x] / (1 - H\_diag[x]))  
norm\_prob\_plot(residual\_var = PRESS\_res, x\_label = 'PRESS Residuals')  
res\_vs\_fitted\_plot(residual\_var = PRESS\_res,  
 main\_title = 'PRESS Residuals vs. Predicted Response',  
 y\_label = 'PRESS Residuals')  
  
par(mfrow = c(2,2))  
plot(X[,2], e,  
 main = 'Residuals vs. log Weeks',  
 ylab = 'Residuals',  
 xlab = 'log Weeks')  
abline(h = 0)  
  
plot(X[,2], studentized\_residuals,  
 main = 'Studentized Residuals vs. log Weeks',  
 ylab = 'Studentized Residuals',  
 xlab = 'log Weeks')  
abline(h = 0)  
  
plot(X[,2], R\_student\_res,  
 main = 'R-Student vs. log Weeks',  
 ylab = 'R-Student',  
 xlab = 'log Weeks')  
abline(h = 0)  
  
plot(X[,2], standardized\_res,  
 main = 'Standardized Residuals vs. log Weeks',  
 ylab = 'Standardized Residuals',  
 xlab = 'log Weeks')  
abline(h = 0)  
  
par(mfrow = c(1,1))  
plot(X[,2], PRESS\_res,  
 main = 'PRESS Residuals vs. log Weeks',  
 ylab = 'PRESS Residuals',  
 xlab = 'log Weeks')  
abline(h = 0)