homework5.R

Barbara Cernosa

2025-03-31

library(foreign)  
library(psych)  
library(car)

library(lattice)  
  
# Chosen dataset for the analysis  
data("USArrests")  
head(USArrests)

## Murder Assault UrbanPop Rape  
## Alabama 13.2 236 58 21.2  
## Alaska 10.0 263 48 44.5  
## Arizona 8.1 294 80 31.0  
## Arkansas 8.8 190 50 19.5  
## California 9.0 276 91 40.6  
## Colorado 7.9 204 78 38.7

us\_crime\_data <- USArrests  
str(us\_crime\_data)

## 'data.frame': 50 obs. of 4 variables:  
## $ Murder : num 13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...  
## $ Assault : int 236 263 294 190 276 204 110 238 335 211 ...  
## $ UrbanPop: int 58 48 80 50 91 78 77 72 80 60 ...  
## $ Rape : num 21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...

psych::describe(us\_crime\_data)

## vars n mean sd median trimmed mad min max range skew  
## Murder 1 50 7.79 4.36 7.25 7.53 5.41 0.8 17.4 16.6 0.37  
## Assault 2 50 170.76 83.34 159.00 168.48 110.45 45.0 337.0 292.0 0.22  
## UrbanPop 3 50 65.54 14.47 66.00 65.88 17.79 32.0 91.0 59.0 -0.21  
## Rape 4 50 21.23 9.37 20.10 20.36 8.60 7.3 46.0 38.7 0.75  
## kurtosis se  
## Murder -0.95 0.62  
## Assault -1.15 11.79  
## UrbanPop -0.87 2.05  
## Rape 0.08 1.32

####### 1. Analysis and model fitting ########################################  
  
########## Correlation between variables  
cor(us\_crime\_data)

## Murder Assault UrbanPop Rape  
## Murder 1.00000000 0.8018733 0.06957262 0.5635788  
## Assault 0.80187331 1.0000000 0.25887170 0.6652412  
## UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412  
## Rape 0.56357883 0.6652412 0.41134124 1.0000000

#Murder and Assault: There’s a strong positive correlation, meaning as assault rates increase, murder rates tend to increase as well.  
#Murder and UrbanPop: Weak correlation - urban population size has little to no relationship with murder rates.  
#Murder and Rape: Moderate correlation - indicating some connection between murder and rape rates.  
  
  
########## Linear regression models:  
# For Murder   
fit1 <- lm(Murder ~ ., data = us\_crime\_data)  
summary(fit1)

##   
## Call:  
## lm(formula = Murder ~ ., data = us\_crime\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3990 -1.9127 -0.3444 1.2557 7.4279   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.276639 1.737997 1.885 0.0657 .   
## Assault 0.039777 0.005912 6.729 2.33e-08 \*\*\*  
## UrbanPop -0.054694 0.027880 -1.962 0.0559 .   
## Rape 0.061399 0.055740 1.102 0.2764   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.574 on 46 degrees of freedom  
## Multiple R-squared: 0.6721, Adjusted R-squared: 0.6507   
## F-statistic: 31.42 on 3 and 46 DF, p-value: 3.322e-11

# This model shows that Assault has a significant positive relationship with Murder while UrbanPop has a small negative relationship.  
# This model explains 67.21% of the variance in Murder (R^2). F-statistic and the p-value indicate that this model is statistically significant.  
  
target <- "Murder"  
predictors <- c("Assault", "UrbanPop", "Rape")  
data <- na.omit(us\_crime\_data[c(target, predictors)])  
n <- nrow(data)  
  
# Total sum of squares (SStotal) for Murder  
SS\_total\_murder <- sum((data$Murder - mean(data$Murder))^2)  
  
# Linear regression model for Murder  
fit\_lm <- lm(Murder ~ ., data = data)  
summary(fit\_lm)

##   
## Call:  
## lm(formula = Murder ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3990 -1.9127 -0.3444 1.2557 7.4279   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.276639 1.737997 1.885 0.0657 .   
## Assault 0.039777 0.005912 6.729 2.33e-08 \*\*\*  
## UrbanPop -0.054694 0.027880 -1.962 0.0559 .   
## Rape 0.061399 0.055740 1.102 0.2764   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.574 on 46 degrees of freedom  
## Multiple R-squared: 0.6721, Adjusted R-squared: 0.6507   
## F-statistic: 31.42 on 3 and 46 DF, p-value: 3.322e-11

# This model shows that Assault has a significant positive relationship with Murder while UrbanPop has a small negative relationship.  
# This model explains 67.21% of the variance in Murder (R^2). F-statistic and the p-value indicate that this model is statistically significant.  
lm\_pred <- predict(fit\_lm, newdata = data)  
R2\_lm\_orig <- 1 - sum((lm\_pred - data$Murder)^2) / SS\_total\_murder  
  
################ 2. Validations ##########################################  
# 10-fold cross-validation  
k <- 10  
set.seed(2010)  
part <- rep(1:k, length.out = n)  
part <- sample(part)  
  
lm\_pred\_cv <- rep(NA, n)  
  
for (i in 1:k) {  
 fit\_lm\_cv <- lm(Murder ~ ., data = data[part != i, ])  
 lm\_pred\_cv[part == i] <- predict(fit\_lm\_cv, newdata = data[part == i, ])  
}  
  
# R^2 for cross-validation  
R2\_lm\_cv <- 1 - sum((lm\_pred\_cv - data$Murder)^2) / SS\_total\_murder  
R2\_lm\_orig

## [1] 0.6720656

R2\_lm\_cv

## [1] 0.6202724

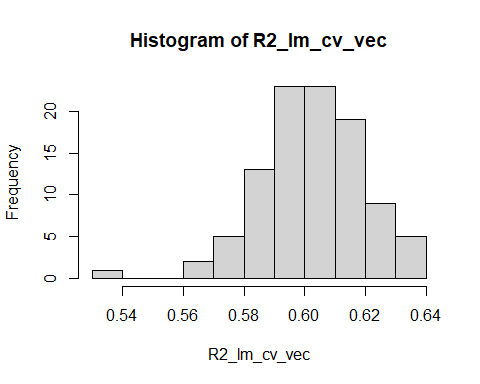
#Low R2 - the model may not generalize perfectly, but it’s still doing a   
# good job of predicting Murder with 60.19% of the variation explained on new data.   
  
# Jackknife resampling  
lm\_pred\_jack <- rep(NA, n)  
  
for (i in 1:n) {  
 fit\_lm\_jack <- lm(Murder ~ ., data = data[-i, ])  
 lm\_pred\_jack[i] <- predict(fit\_lm\_jack, newdata = data[i, ])  
}  
  
# R^2 for jackknife resampling  
R2\_lm\_jack <- 1 - sum((lm\_pred\_jack - data$Murder)^2) / SS\_total\_murder  
R2\_lm\_orig

## [1] 0.6720656

R2\_lm\_jack

## [1] 0.6054642

# 100 repetitions of 10-fold cross-validation  
m <- 100  
R2\_lm\_cv\_vec <- rep(NA, m)  
  
for (j in 1:m) {  
 part <- sample(part)  
   
 lm\_pred\_cv <- rep(NA, n)  
 for (i in 1:k) {  
 fit\_lm\_cv <- lm(Murder ~ ., data = data[part != i, ])  
 lm\_pred\_cv[part == i] <- predict(fit\_lm\_cv, newdata = data[part == i, ])  
 }  
   
 R2\_lm\_cv\_vec[j] <- 1 - sum((lm\_pred\_cv - data$Murder)^2) / SS\_total\_murder  
}  
  
R2\_lm\_cv\_mean <- mean(R2\_lm\_cv\_vec)  
hist(R2\_lm\_cv\_vec)



R2\_lm\_orig

## [1] 0.6720656

R2\_lm\_cv\_mean

## [1] 0.6019301

# Bootstrap resampling  
m <- 1000  
diffErrSsBoot\_lm <- numeric(m)  
  
set.seed(2010)  
for (i in 1:m) {  
 boot\_ids <- sample(1:n, size = n, replace = TRUE)  
 fit\_lm\_boot <- lm(Murder ~ ., data = data[boot\_ids, ])  
 errSsBoot\_lm <- sum(fit\_lm\_boot$resid^2)  
 predOrg\_lm <- predict(fit\_lm\_boot, newdata = data)  
 errSsOrg\_lm <- sum((predOrg\_lm - data$Murder)^2)  
 diffErrSsBoot\_lm[i] <- errSsOrg\_lm - errSsBoot\_lm  
}  
  
mean(diffErrSsBoot\_lm)

## [1] 54.79012

# Calculated R^2 from bootstrap  
R2\_lm\_boot <- 1 - (sum(fit\_lm$resid^2) + mean(diffErrSsBoot\_lm)) / SS\_total\_murder  
R2\_lm\_orig

## [1] 0.6720656

R2\_lm\_boot

## [1] 0.6131232

# Comparing R2 values  
R2s <- c(R2\_lm\_orig = R2\_lm\_orig, R2\_lm\_cv = R2\_lm\_cv,   
 R2\_lm\_cv\_mean = R2\_lm\_cv\_mean, R2\_lm\_jack = R2\_lm\_jack,   
 R2\_lm\_boot = R2\_lm\_boot)  
R2s

## R2\_lm\_orig R2\_lm\_cv R2\_lm\_cv\_mean R2\_lm\_jack R2\_lm\_boot   
## 0.6720656 0.6202724 0.6019301 0.6054642 0.6131232

# Interpretation:  
# The original R² was 0.67, showing that Linear Regression explains 67.2% of the variance in Murder rates. The 10-fold CV R² was 0.620, lower than the original, suggesting the model doesn’t generalize perfectly but still explains about 62% of the variance on new data. Jackknife R² was 0.605, close to 10-fold CV, indicating consistent performance. The 100-repetition CV mean R² was 0.602, and bootstrap R² was 0.613. All these R² values are fairly close to each other, indicating that the model generalizes well across different validations. The small variations suggest that the model is quite stable and works similarly in different testing conditions.  
  
  
############# 2. Predictions ################################################  
crime\_pred\_murder\_lm <- predict(fit\_lm, newdata = data)  
crime\_predictions <- data.frame(  
 Murder\_Predicted\_LM = crime\_pred\_murder\_lm  
)  
head(crime\_predictions)

## Murder\_Predicted\_LM  
## Alabama 10.793487  
## Alaska 13.845014  
## Arizona 12.499018  
## Arkansas 9.296908  
## California 11.770833  
## Colorado 9.501235

head(data)

## Murder Assault UrbanPop Rape  
## Alabama 13.2 236 58 21.2  
## Alaska 10.0 263 48 44.5  
## Arizona 8.1 294 80 31.0  
## Arkansas 8.8 190 50 19.5  
## California 9.0 276 91 40.6  
## Colorado 7.9 204 78 38.7

# Comparing predictions to actual Murder rates for the first few states, Alabama’s predicted rate was 10.79 vs. actual 13.2 (underpredicted by 2.41), while Alaska’s was 13.85 vs. 10.0 (overpredicted by 3.85). These differences vary, with errors ranging from -2.41 to +4.40, but the R² values capture the overall fit across all states.