STOR 455 Cross Validation Correlation Practice

The data in FirstYearGPA contains information on 219 college students.

Variable Description HSGPA | High school GPA SATV | Verbal/critical reading SAT score SATM | Math SAT score Male | 1 for male, 0 for female HU | Number of credit hours earned in humanities courses in high school SS | Number of credit hours earned in social science courses in high school FirstGen | 1 if the student is the first in her or his family to attend college White | 1 for white students, 0 for others CollegeBound | 1 if attended a high school where of students intend to go on to college

library(Stat2Data)  
data("FirstYearGPA")

#Creating training sample  
GPATrain = data.frame(FirstYearGPA[c(0:150),])  
  
#Creating holdout sample  
GPAHoldout = data.frame(FirstYearGPA[c(150:219),])

1. Use the training sample to fit a multiple regression to predict GPA using HSGPA, HU, and White. Compute the predicted GPA for each case in the holdout sample using this model, then compute the residuals for each of the holdout cases.

GPAModel = lm(GPA~HSGPA + HU + factor(White), data=GPATrain)  
fitGPA = predict(GPAModel, newdata=GPAHoldout)  
GPAResid = GPAHoldout$GPA-fitGPA  
GPAResid

## 150 151 152 153 154   
## -0.6289464013 0.1774394377 0.7093971450 0.7199727824 -0.5703145067   
## 155 156 157 158 159   
## -0.4140782175 -0.0382432643 -0.2047372186 -0.6550454987 0.3798684276   
## 160 161 162 163 164   
## 0.4580618768 0.1985923190 0.2664629939 -0.2747119553 0.2351734504   
## 165 166 167 168 169   
## -0.7751471876 -0.2536553431 -0.0018953044 -0.3524895565 -0.3600627205   
## 170 171 172 173 174   
## 0.0141157790 0.7683507724 -0.7868886116 0.2257948302 0.6026659164   
## 175 176 177 178 179   
## -0.0881825223 -0.6084714179 0.4681877699 0.2216631324 -0.0818348277   
## 180 181 182 183 184   
## 0.1975249313 -0.3461618179 -0.6202984124 0.2034320264 -0.4424330626   
## 185 186 187 188 189   
## 0.1059798830 -0.4148954015 -0.4579081385 -0.3659294161 -0.3708417600   
## 190 191 192 193 194   
## -0.0751703326 0.1907910433 -0.5770129223 -0.1300272176 0.0181296694   
## 195 196 197 198 199   
## 0.4429532055 -0.5626864554 0.3596999810 0.2423307665 -0.2842874743   
## 200 201 202 203 204   
## -0.1221447143 0.1743707542 -0.6322232584 -0.5838442333 0.7272796483   
## 205 206 207 208 209   
## -0.7945570625 0.1053929468 -0.1978247748 -0.4516368696 0.4624243698   
## 210 211 212 213 214   
## 0.4690425865 0.4176008337 -0.2202564063 0.0842287668 0.0529403002   
## 215 216 217 218 219   
## 0.0408244100 -0.2425121782 -0.2950094662 -0.0006948518 -0.1901644095

1. Compute the mean and standard deviation for the residuals. Is the mean reasonably close to zero? Is the standard deviation reasonably close to the standard deviation of the error term from the fit to the training sample?

mean(GPAResid)

## [1] -0.06760761

sd(GPAResid)

## [1] 0.4092978

summary(GPAModel)

##   
## Call:  
## lm(formula = GPA ~ HSGPA + HU + factor(White), data = GPATrain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.09844 -0.23079 0.03517 0.23600 0.82933   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.147478 0.311524 3.683 0.000323 \*\*\*  
## HSGPA 0.466053 0.088393 5.273 4.75e-07 \*\*\*  
## HU 0.015328 0.004091 3.747 0.000257 \*\*\*  
## factor(White)1 0.199174 0.076152 2.615 0.009846 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3773 on 146 degrees of freedom  
## Multiple R-squared: 0.2842, Adjusted R-squared: 0.2695   
## F-statistic: 19.32 on 3 and 146 DF, p-value: 1.319e-10

The mean of -0.06760761 is reasonably close to 0. The standard deviation of the residuals of 0.4092978 is quite close to the standard deviation from our model of 0.3773. This points to our model being a good predictor of GPA in general, not just for our specific training sample.

1. Compute the cross-validation correlation between the actual and predicted GPA values for the cases in the holdout sample.

crosscor = cor(GPAHoldout$GPA, fitGPA)  
crosscor

## [1] 0.5795404

1. Square the cross-validation correlation and subtract from for the training sample to compute the shrinkage. Does it look like the training model works reasonably well for the holdout sample or has there been a considerable drop in the amount of variability explained?

crosscor^2

## [1] 0.3358671

summary(GPAModel)$r.squared - crosscor^2

## [1] -0.05166974

The shrinkage here is actually a negative value, which means that my model created from the training data actually predicts GPA from the holdout model better than the original training data. The values in the holdout model are predicted around 5% better than the training model values, which is not a considerable problem as it is less than 10% and does not point to a significant difference in the effectiveness of model prediction between two datasets.