Lab and HW 19

Vy Duong, Vi Ly, Shawn Olichwier

4/23/2019

From HW and Lab 18, validate your best two models and compare.

```
dat=read.csv("/Users/vyduong/Downloads/B13 copy.csv",header=T)
data.frame(dat)
```

```
##       y   x1    x2    x3  x4   x5  x6
## 1  4540 2140 20640 30250 205 1732  99
## 2  4315 2016 20280 30010 195 1697 100
## 3  4095 1905 19860 29780 184 1662  97
## 4  3650 1675 18980 29330 164 1598  97
## 5  3200 1474 18100 28960 144 1541  97
## 6  4833 2239 20740 30083 216 1709  87
## 7  4617 2120 20305 29831 206 1669  87
## 8  4340 1990 19961 29604 196 1640  87
## 9  3820 1702 18916 29088 171 1572  85
## 10 3368 1487 18012 28675 149 1522  85
## 11 4445 2107 20520 30120 195 1740 101
## 12 4188 1973 20130 29920 190 1711 100
## 13 3981 1864 19780 29720 180 1682 100
## 14 3622 1674 19020 29370 161 1630 100
## 15 3125 1440 18030 28940 139 1572 101
## 16 4560 2165 20680 30160 208 1704  98
## 17 4340 2048 20340 29960 199 1679  96
## 18 4115 1916 19860 29710 187 1642  94
## 19 3630 1658 18950 29250 164 1576  94
## 20 3210 1489 18700 28890 145 1528  94
## 21 4330 2062 20500 30190 193 1748 101
## 22 4119 1929 20050 29960 183 1713 100
## 23 3891 1815 19680 29770 173 1684 100
## 24 3467 1595 18890 29360 153 1624  99
## 25 3045 1400 17870 28960 134 1569 100
## 26 4411 2047 20540 30160 193 1746  99
## 27 4203 1935 20160 29940 184 1714  99
## 28 3968 1807 19750 29760 173 1679  99
## 29 3531 1591 18890 29350 153 1621  99
## 30 3074 1388 17870 28910 133 1561  99
## 31 4350 2071 20460 30180 198 1729 102
## 32 4128 1944 20010 29940 186 1692 101
## 33 3940 1831 19640 29750 178 1667 101
## 34 3480 1612 18710 29360 156 1609 101
## 35 3064 1410 17780 28900 136 1552 101
## 36 4402 2066 20520 30170 197 1758 100
```

```
## 37 4180 1954 20150 29950 188 1729  99
## 38 3973 1835 19750 29740 178 1690  99
## 39 3530 1616 18850 29320 156 1616  99
## 40 3080 1407 17910 28910 137 1569 100
```

```
m1=lm(y~x1+x6,data=dat)   # Model 1
m2=lm(y~x4,data=dat)      # Model 2

#inspection and comparison of coefficients
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x6, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.586 -18.442  -1.569  15.451  82.512
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1103.27736  121.94030   9.048 6.52e-11 ***
## x1             1.98197    0.02163  91.637  < 2e-16 ***
## x6            -8.07277    1.15524  -6.988 2.94e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.94 on 37 degrees of freedom
## Multiple R-squared:  0.9957, Adjusted R-squared:  0.9955
## F-statistic:  4292 on 2 and 37 DF,  p-value: < 2.2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = y ~ x4, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -88.96 -31.51 -12.91  25.16 110.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 164.9901    60.7182   2.717  0.00986 **
## x4           21.4270     0.3449  62.119  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.49 on 38 degrees of freedom
## Multiple R-squared:  0.9902, Adjusted R-squared:   0.99
## F-statistic:  3859 on 1 and 38 DF,  p-value: < 2.2e-16
```

- We used all possible regressions to develop the best two models for the table B.13
- Model 1: $\hat{y} = 1103.27736 + 1.98197 * x1 - 8.07277 * x6$
- Model 2: $\hat{y} = 164.9901 + 21.4270 * x4$
- Model 1 contains x1 and x6 while model 2 contains only x4. We will calculate the values of the PRESS statistics, R-squared prediction, and the VIF's for both models.

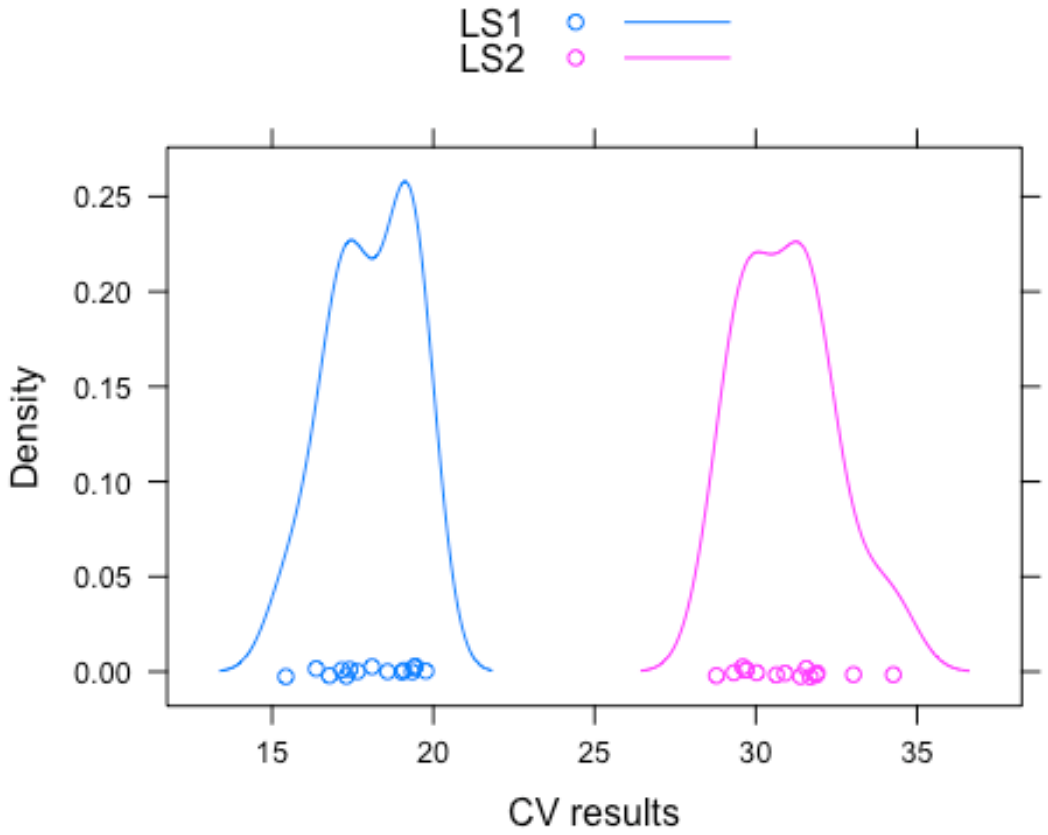
```
# VIF's
library(car)

## Loading required package: carData

vif(m1)

##       x1       x6
## 1.00532 1.00532
```

- For model 1, both VIFs are smaller than 5, indicating no potential problems with multicollinearity. However, for model 2, we cannot find the VIFs. The model contains only 1 term (x4).

```
library(MPV)
anova1 = anova(m1)
sst1 = sum(anova1$'Sum Sq') #Calculate the total sum of squares
PRESS(m1)

## [1] 48396.43

1-PRESS(m1)/(sst1)   # Calculate the predictive R^2

## [1] 0.9951272

anova2 = anova(m2)
sst2 = sum(anova2$'Sum Sq') #Calculate the total sum of squares
PRESS(m2)

## [1] 106716.6

1-PRESS(m2)/(sst2)   # Calculate the predictive R^2

## [1] 0.9892553
```

- For model 1, the PRESS statistic is 48396.43 and the predictive R-squared is 0.9951272
- For model 2, the PRESS statistic is 106716.6 and the predictive R-squared is 0.9892553
- According to the above result, model 1 have a larger value of the predictive R-square. It means model 1 is better than model 2.

```r
library(cvTools)

## Loading required package: lattice

## Loading required package: robustbase

## Warning: package 'robustbase' was built under R version 3.5.2

folds <- cvFolds(nrow(dat), K = 4, R =15) #type = "random", "consecutive",
"interleaved"
cvfit1 <- cvLm(m1, cost = rtmspe,folds = folds)
cvfit2 <- cvLm(m2, cost = rtmspe,folds = folds)
cvFits <- cvSelect(LS1 = cvfit1, LS2 =cvfit2)
cvFits

##
## 4-fold CV results:
##   Fit       CV
## 1 LS1 18.05945
## 2 LS2 30.95810
##
## Best model:
##    CV
## "LS1"

densityplot(cvFits) #plot combined results
```



- We split the data by running the cvFolds, with K = 4 (split the observations into 4 groups) and R = 15 (repeat K-fold cross-validation by 15). It gives us the result with model 1 is the best model.
- According to the regression models, the PRESS statistics, the R-squared predictions, the VIF's, and the data splitting for both models, model 1 is the best model.