Name - vvgupta2

Homework 12

Homework Instructions

For questions that require code, please create a code chunk directly below the question and type your code there. Your knitted pdf will show both your code and your output. You are encouraged to knit your file as you work to check that your coding and formatting is done so appropriately.

For written responses or multiple choice questions, please bold your (selected) answer.

Grading Details

All questions will be graded full credit (1 point), half credit (0.5 point) or no credit (0 points).

Full credit responses should have the correct response and appropriate code (if applicable). Half credit responses will have a reasonable attempt (typically no more than one small error or oversight), and no credit responses will be either non-attempts or attempts with significant errors.

Exercise 1

For exercises 1 - 8, use the the built-in `R` dataset `mtcars`. Use `mpg` as the response variable. Do not modify any of the data. (An argument could be made for `cyl`, `gear`, and `carb` to be coerced to factors, but for simplicity, we will keep them numeric.)

```
# starter
mtcars
```

Fit an additive linear model with all available variables as predictors. Call this model `full_mod`. Which predictor seems to be most explained by the other predictors in the model?

Hint: We learned a function in chapter 15 that can help here!

```
# solution
full_mod = lm(mpg ~ ., data = mtcars)
summary(full_mod)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
```

```
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```r
library(faraway)
vif(full_mod)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs        am
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487
##      gear      carb
##  5.357452  7.908747
```

Exercise 2

Create a function that takes a model name as input and calculates the LOOCV-RMSE of that model.

Once created, calculate the LOOCV-RMSE of the model fit in Exercise 1.

```r
# solution
calc_loocv_rmse = function(model) {
  sqrt(mean((resid(model) / (1 - hatvalues(model))) ^ 2))
}
calc_loocv_rmse(full_mod)
```

```
## [1] 3.490209
```

Exercise 3

Start with the full model, and then perform variable selection using backwards AIC. Call this model `selected`. Then print the model coefficients of the selected model.

Tip: use `trace = 0` as an argument to suppress lengthy output

```r
# solution
selected = step(full_mod, direction = "backward", trace = 0)
coef(selected)
```

```
## (Intercept)          wt        qsec          am
##    9.617781   -3.916504    1.225886    2.935837
```

Exercise 4

What is the LOOCV-RMSE of the "selected" model?

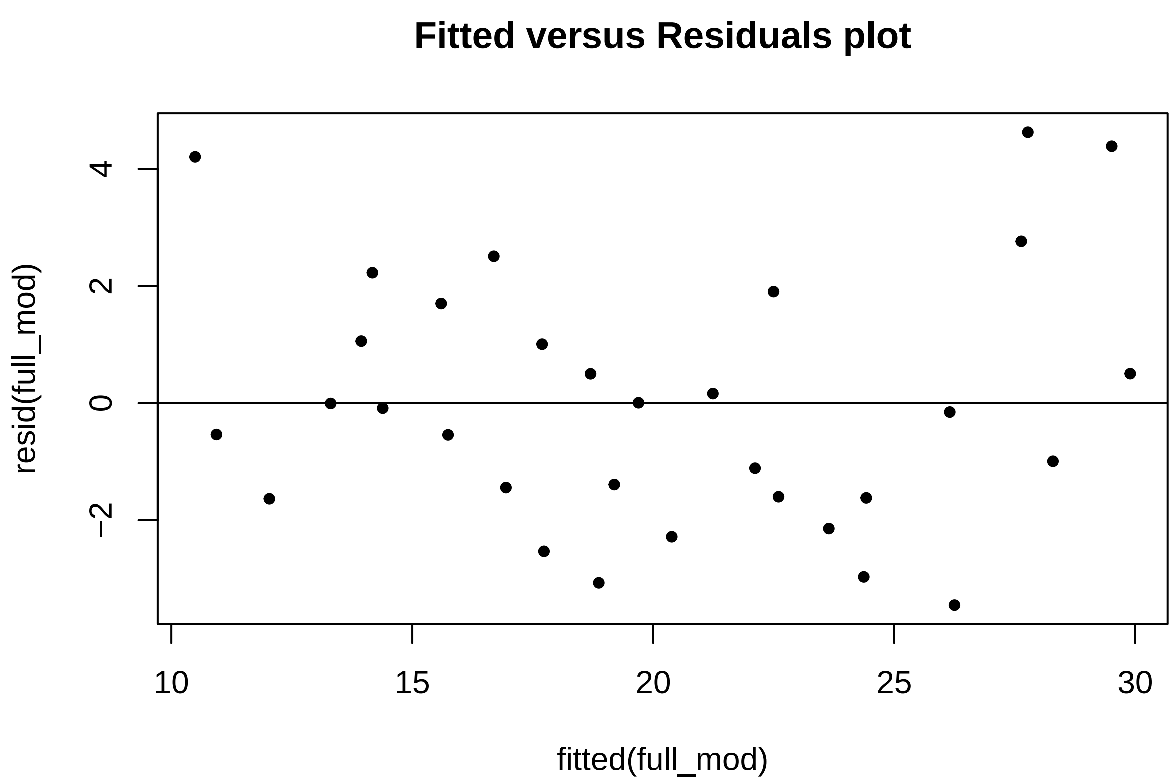
```r
# solution
calc_loocv_rmse(selected)
```

```
## [1] 2.688538
```

Exercise 5

Create a residual plot for the **full model**, and also run a Breusch-Pagan test and Shapiro-Wilk test.

```r
# solution
plot (fitted(full_mod), resid (full_mod), main = "Fitted versus Residuals plot", pch=20)
abline(h=0)
```




```r
shapiro.test(resid (full_mod))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(full_mod)
## W = 0.95694, p-value = 0.2261
```

```
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

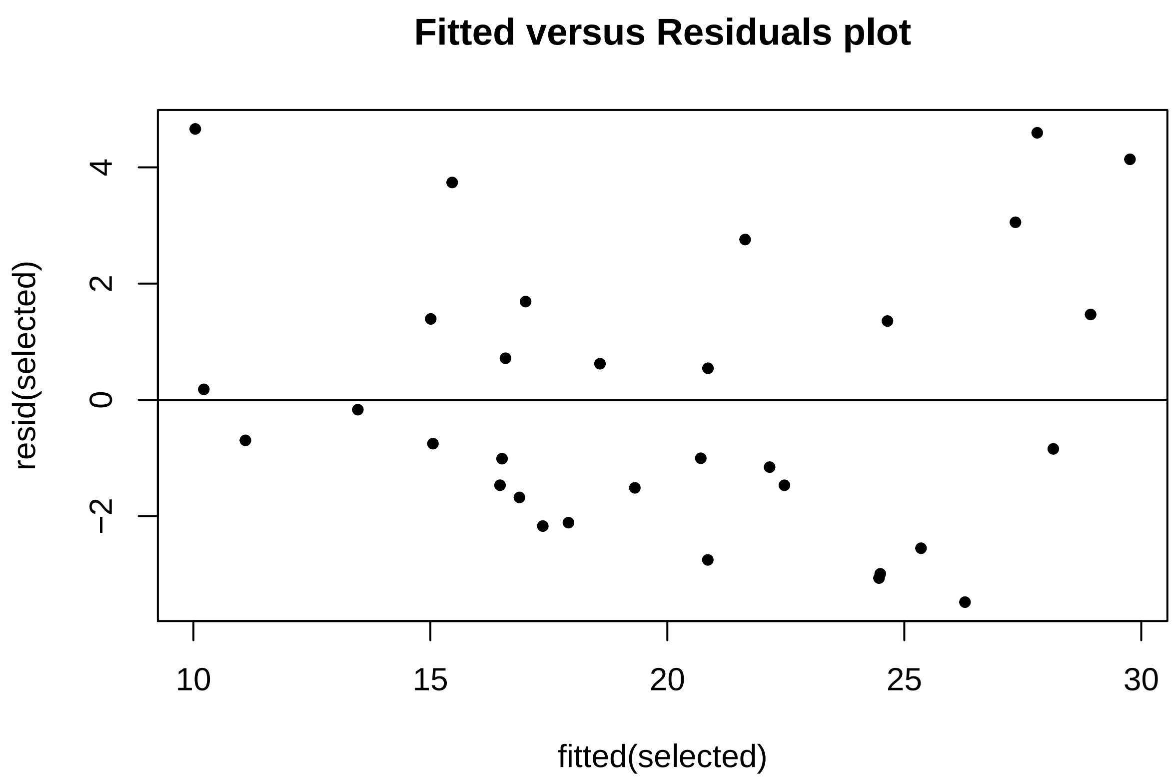
```
bptest (full_mod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  full_mod
## BP = 14.914, df = 10, p-value = 0.1352
```

Exercise 6

Create a residual plot for the **selected model**, and also run a Breusch-Pagan test and Shapiro-Wilk test.

```
# solution
plot (fitted(selected), resid (selected), main = "Fitted versus Residuals plot", pch=20)
abline(h=0)
```




```
shapiro.test(resid (selected))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(selected)
## W = 0.9411, p-value = 0.08043
```

```
bptest (selected)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  selected
## BP = 6.1871, df = 3, p-value = 0.1029
```

Exercise 7

Based on the previous exercises, which model is probably better for predicting?

- The selected model: It has better accuracy, and the model diagnostics are reasonable
- The selected model: The model diagnostics for the full model are problematic and will lead to unreliable estimates

- The full model: It has better accuracy, and the model diagnostics are reasonable

- The full model: The model diagnostics for the selected model are problematic and will lead to unreliable estimates

Exercise 8

```
# starter
LifeCycleSavings
```

For exercises 8 - 12, use the the built-in `R` dataset `LifeCycleSavings`. Use `sr` as the response variable.

Fit a model with all available predictors as well as their two-way interactions. What is the Adjusted R^2 of this model?

Hint: You can use the "^2" notation to do this concisely–see 16.3 from the book for an example.

```
# solution
model1 = lm(sr ~ .^2, data = LifeCycleSavings)
summary(model1)$adj.r.squared
```

```
## [1] 0.261233
```

Exercise 9

Start with the model fit in Exercise 8, then perform variable selection using backwards **BIC**. Print the coefficients of the selected model.

```
# solution
n = length(resid(model1))
step(model1, direction = "backward", k = log(n), trace = 0)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + dpi + ddpi + dpi:ddpi, data = LifeCycleSavings)
##
## Coefficients:
## (Intercept)        pop15          dpi         ddpi     dpi:ddpi
##  16.5287997   -0.2023669   -0.0027411    0.0462479    0.0008171
```

```
back_bic = step(model1, direction = "backward", k = log(n), trace = 0)
coef(back_bic)
```

```
##   (Intercept)         pop15           dpi          ddpi      dpi:ddpi
## 16.5287997486 -0.2023668518 -0.0027411096  0.0462478987  0.0008170652
```

Exercise 10

Start with the model fit in Exercise 8, then perform variable selection using backwards AIC. Print the coefficients of the selected model.

```r
# solution
step(model1, direction = "backward")
```

```
## Start:  AIC=144.41
## sr ~ (pop15 + pop75 + dpi + ddpi)^2
##
##                Df Sum of Sq    RSS    AIC
## - pop15:dpi    1    0.0003 578.37 142.41
## - pop15:ddpi   1    0.4910 578.86 142.45
## - pop75:dpi    1    5.5364 583.91 142.89
## - pop75:ddpi   1    6.5316 584.90 142.97
## - pop15:pop75  1    8.0062 586.38 143.10
## - dpi:ddpi     1   16.2285 594.60 143.79
## <none>                     578.37 144.41
##
## Step:  AIC=142.41
## sr ~ pop15 + pop75 + dpi + ddpi + pop15:pop75 + pop15:ddpi +
##     pop75:dpi + pop75:ddpi + dpi:ddpi
##
##                Df Sum of Sq    RSS    AIC
## - pop15:ddpi   1    0.5108 578.88 140.45
## - pop75:ddpi   1    6.7381 585.11 140.99
## - pop75:dpi    1    8.3861 586.76 141.13
## - pop15:pop75  1    9.5161 587.89 141.23
## - dpi:ddpi     1   16.4076 594.78 141.81
## <none>                     578.37 142.41
##
## Step:  AIC=140.45
## sr ~ pop15 + pop75 + dpi + ddpi + pop15:pop75 + pop75:dpi + pop75:ddpi +
##     dpi:ddpi
##
##                Df Sum of Sq    RSS    AIC
## - pop75:ddpi   1     6.546 585.43 139.02
## - pop15:pop75  1     9.910 588.79 139.30
## - pop75:dpi    1    10.734 589.62 139.37
## <none>                     578.88 140.45
## - dpi:ddpi     1    44.067 622.95 142.12
##
## Step:  AIC=139.02
## sr ~ pop15 + pop75 + dpi + ddpi + pop15:pop75 + pop75:dpi + dpi:ddpi
##
##                Df Sum of Sq    RSS    AIC
## - pop75:dpi    1    10.168 595.60 137.88
## - pop15:pop75  1    12.249 597.68 138.05
## <none>                     585.43 139.02
## - dpi:ddpi     1    41.092 626.52 140.41
##
## Step:  AIC=137.88
## sr ~ pop15 + pop75 + dpi + ddpi + pop15:pop75 + dpi:ddpi
##
##                Df Sum of Sq    RSS    AIC
## - pop15:pop75  1     2.958 598.55 136.12
## <none>                     595.60 137.88
```

```
## - dpi:ddpi     1    45.223 640.82 139.54
##
## Step:  AIC=136.12
## sr ~ pop15 + pop75 + dpi + ddpi + dpi:ddpi
##
##            Df Sum of Sq    RSS    AIC
## - pop75     1    16.708 615.26 135.50
## <none>                  598.55 136.12
## - dpi:ddpi  1    52.158 650.71 138.30
## - pop15     1    63.923 662.48 139.20
##
## Step:  AIC=135.5
## sr ~ pop15 + dpi + ddpi + dpi:ddpi
##
##            Df Sum of Sq    RSS    AIC
## <none>                  615.26 135.50
## - pop15     1    58.171 673.43 138.02
## - dpi:ddpi  1    70.686 685.95 138.94
```

```
##
## Call:
## lm(formula = sr ~ pop15 + dpi + ddpi + dpi:ddpi, data = LifeCycleSavings)
##
## Coefficients:
## (Intercept)        pop15          dpi         ddpi     dpi:ddpi
##  16.5287997   -0.2023669   -0.0027411    0.0462479    0.0008171
```

```
#note we can also turn off with trace = 0
back_aic_2 = step(model1, direction = "backward", trace = 0)
coef(back_aic_2)
```

```
##   (Intercept)         pop15           dpi          ddpi      dpi:ddpi
## 16.5287997486 -0.2023668518 -0.0027411096  0.0462478987  0.0008170652
```

If done correctly, you'll find that AIC and BIC in this case actually produce the same model! It doesn't always happen that way, but sometimes it does.

Exercise 11

Consider the full model in Exercise 8 and the BIC model (or AIC) model fit next. Based on LOOCV-RMSE, which of these models is likely more accurate at prediction?

```
# solution
calc_loocv_rmse(back_bic)
```

```
## [1] 3.833628
```

```
calc_loocv_rmse(back_aic_2)
```

```
## [1] 3.833628
```

- The full model is likely more accurate at prediction
- **The selected model is likely more accurate at prediction**

Exercise 12

Which **one** of these statements best distinguishes AIC vs. BIC for model selection? Bold your answer.

- AIC will generally favor more interaction terms and fewer solo predictors
- BIC will generally favor more interaction terms and fewer solo predictors
- AIC will generally favor smaller models with fewer parameters

- BIC will generally favor smaller models with fewer parameters