SDS 272: Homework #6

My Name

For homework assignments, you only need to submit your knitted pdf file to Moodle, but be sure your RMarkdown file is saved and accessible in your Submit folder on the RStudio server. Include all of your R code, in addition to your output, plots, and written responses. Read each question carefully, and be sure your written responses are thorough, succinct, and clear, with careful use of statistical language. Finally, check your assignment prior to submission; don't just assume it knitted okay.

1) Fertility measurements (described on page 189 of STAT2 - Exercise 4.4) *32 points total*

The data set `Fertility` has been loaded in the setup R chunk (you can check it out in the chunk below).

a) Create a correlation matrix and comment on what you learn about the relationships between numeric variables.

We see that age is positively related to FSH, MaxDailyGn, and TotalGn, and it is inversely related to Mean AFC, E2, MaxE2, and Oocytes. MeanAFC is positively related to E2, MaxE2, and Oocytes, and inversely related to FSH, MaxDailyGn, and TotalGn. FSH is positively correlated with MaxDailyGn and TotalGn, and inversely correlated with E2, MaxE2, and Oocytes. E2 is correlated with MaxE2 and Oocytes, and inversely correlated with MaxDailyGn and Total GN. MaxE2 has the same relationships as E2. MaxDailyGn is positively related with TotalGn and inversely related to Oocytes. Total Gn is negatively related to Oocytes.

b) Fit the multiple regression model to predict Average antral follical count (`MeanAFC`) using `Age`, `FSH`, `E2`, `MaxE2`, `MaxDailyGn`, `TotalGn`, `Oocytes`, and `Embryos` as predictors. Report on which variable(s) *seem* to be most important for the model based on the summary of this model.

It looks like Oocytes, E2, and MaxDailyGn seem to be the 3 most important.

```r
library(vip)
```

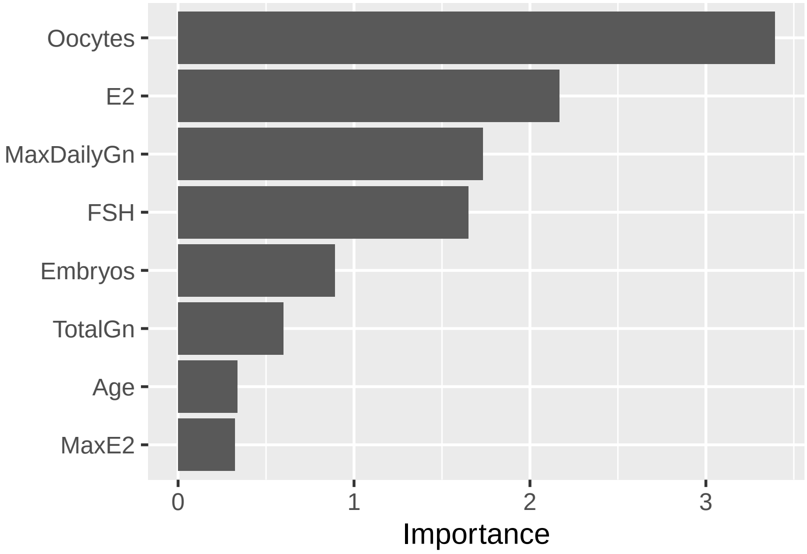
```
##
## Attaching package: 'vip'

## The following object is masked from 'package:utils':
##
##     vi
```

```r
follical_model<-lm(MeanAFC~Age+FSH+E2+MaxE2+MaxDailyGn+TotalGn+Oocytes+Embryos, data=Fertility)
summary(follical_model)
```

```
##
## Call:
## lm(formula = MeanAFC ~ Age + FSH + E2 + MaxE2 + MaxDailyGn +
##     TotalGn + Oocytes + Embryos, data = Fertility)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.619  -3.545  -0.892   2.674  32.797
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.4430476  3.3105012   5.873 1.06e-08 ***
## Age         -0.0306469  0.0913435  -0.336 0.737456
## FSH         -0.3459110  0.2097527  -1.649 0.100089
## E2          -0.0506100  0.0233484  -2.168 0.030917 *
## MaxE2       -0.0001714  0.0005320  -0.322 0.747484
## MaxDailyGn  -0.0131015  0.0075583  -1.733 0.083975 .
## TotalGn     -0.0003702  0.0006200  -0.597 0.550884
## Oocytes      0.3294810  0.0971744   3.391 0.000784 ***
## Embryos      0.1180175  0.1325688   0.890 0.374000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.383 on 324 degrees of freedom
## Multiple R-squared:  0.2792, Adjusted R-squared:  0.2614
## F-statistic: 15.69 on 8 and 324 DF,  p-value: < 2.2e-16
```

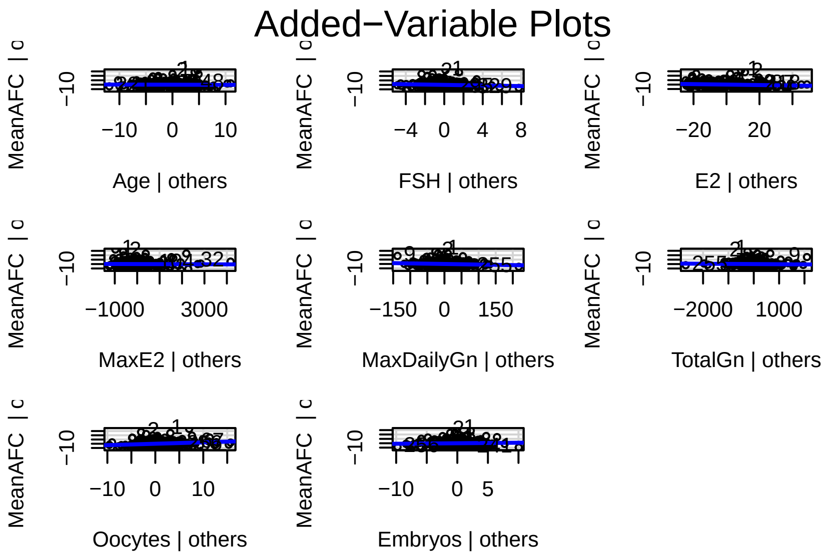
```
vip(follical_model)
```



c) Create the added variable plots for the MLR model in part (b). Comment on how the added variable plots indicate a stronger predictor. Does this method indicate the same important variable(s) you identified from the model summary?

Oocytes seem to have the strongest correlation. This method seems to have similar important variables as the model summary.

```
avPlots(follical_model)
```



d) Find the variance inflation factors for your predictor variables in the MLR model. Do any variable(s) seem to have multicollinearity problems? How is this shown in the added variable plots for those variables?

It seems that there could be multicollinearilty problems in MaxDailyGn and Total Gn. We see that those variables are not strongly correlated on the plots.

```
vif(follical_model)
```

```
##        Age        FSH         E2      MaxE2 MaxDailyGn    TotalGn    Oocytes 
##   1.500520   1.353076   1.029520   1.405840   6.238589   5.892978   2.689422 
##    Embryos 
##   2.385170
```

e) Use the best subsets method to find the 2 best combination of variables for models from one predictor to all eight predictors (there's only one model with all eight). Report the variables included in the overall optimal model using:

```
library(leaps)

subsets <- regsubsets(MeanAFC ~ Age+FSH+E2+MaxE2+MaxDailyGn+TotalGn+Oocytes+Embryos, data = Fertility)

subset_summary <- summary(subsets)
subset_summary$bic
```

```
## [1] -52.07581 -82.64251 -81.03914 -78.56021 -73.56400 -68.12238 -62.44639
## [8] -56.74495
```

```
subset_summary$cp
```

```
## [1] 42.243451  5.827478  3.651309  2.381945  3.589552  5.232483  7.103834
## [8]  9.000000
```

```
subset_summary$adjr2
```

```
## [1] 0.1715932 0.2550655 0.2621768 0.2672894 0.2668385 0.2653986 0.2634306
## [8] 0.2613940
```

```
i) BIC criteria
ii) Mallow's Cp criteria
```

`iii) Adj R^2 criteria`

Are there differences between the “best model” for each criteria?

For BIC, the lowest value is -82.6, meaning that the best model is the one with two predictors: Oocytes and MaxDailyGn.

For Cp, the lowest value is 2.38 for the model with 4 predictors: Oocytes, MaxDailyGn, E2, and FSH.

For Adj Rˆ2 the highest value 0.2673, which is the model with 4 predictors, same as Cp.

f) Using stepwise selection, find the optimal set of variables to predict `MeanAFC`. Report the fitted regression equation and the adjusted Rˆ2 value.

MeanAFC = B0 + B1(FSH) + B2(E2) + B3(MaxDailyGn) + B4(Oocytes)

Adjusted Rˆ2: 0.2673

```r
full_model<-lm(MeanAFC~Age+FSH+E2+MaxE2+MaxDailyGn+TotalGn+Oocytes+Embryos, data=Fertility)


summary(full_model)
```

```
##
## Call:
## lm(formula = MeanAFC ~ Age + FSH + E2 + MaxE2 + MaxDailyGn +
##     TotalGn + Oocytes + Embryos, data = Fertility)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.619  -3.545  -0.892   2.674  32.797
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.4430476  3.3105012   5.873 1.06e-08 ***
## Age         -0.0306469  0.0913435  -0.336 0.737456
## FSH         -0.3459110  0.2097527  -1.649 0.100089
## E2          -0.0506100  0.0233484  -2.168 0.030917 *
## MaxE2       -0.0001714  0.0005320  -0.322 0.747484
## MaxDailyGn  -0.0131015  0.0075583  -1.733 0.083975 .
## TotalGn     -0.0003702  0.0006200  -0.597 0.550884
## Oocytes      0.3294810  0.0971744   3.391 0.000784 ***
## Embryos      0.1180175  0.1325688   0.890 0.374000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.383 on 324 degrees of freedom
## Multiple R-squared:  0.2792, Adjusted R-squared:  0.2614
## F-statistic: 15.69 on 8 and 324 DF,  p-value: < 2.2e-16
```

```r
stepwise_model <- step(full_model, direction = "both")
```

```
## Start:  AIC=1243.44
## MeanAFC ~ Age + FSH + E2 + MaxE2 + MaxDailyGn + TotalGn + Oocytes +
##     Embryos
##
##              Df Sum of Sq   RSS    AIC
## - MaxE2       1      4.23 13206 1241.5
## - Age         1      4.59 13207 1241.5
## - TotalGn     1     14.53 13217 1241.8
## - Embryos     1     32.29 13234 1242.2
## <none>                   13202 1243.4
## - FSH         1    110.82 13313 1244.2
## - MaxDailyGn  1    122.43 13325 1244.5
## - E2          1    191.45 13394 1246.2
## - Oocytes     1    468.44 13671 1253.0
##
## Step:  AIC=1241.54
## MeanAFC ~ Age + FSH + E2 + MaxDailyGn + TotalGn + Oocytes + Embryos
##
##              Df Sum of Sq   RSS    AIC
## - Age         1      5.24 13212 1239.7
## - TotalGn     1     14.41 13221 1239.9
## - Embryos     1     30.37 13237 1240.3
## <none>                   13206 1241.5
## - FSH         1    109.74 13316 1242.3
## - MaxDailyGn  1    119.45 13326 1242.5
## + MaxE2       1      4.23 13202 1243.4
## - E2          1    192.76 13399 1244.4
## - Oocytes     1    477.71 13684 1251.4
##
## Step:  AIC=1239.68
## MeanAFC ~ FSH + E2 + MaxDailyGn + TotalGn + Oocytes + Embryos
##
##              Df Sum of Sq   RSS    AIC
## - TotalGn     1     14.55 13226 1238.0
## - Embryos     1     32.31 13244 1238.5
## <none>                   13212 1239.7
## - FSH         1    111.76 13323 1240.5
## - MaxDailyGn  1    144.07 13356 1241.3
## + Age         1      5.24 13206 1241.5
## + MaxE2       1      4.89 13207 1241.5
## - E2          1    192.68 13404 1242.5
## - Oocytes     1    472.52 13684 1249.4
##
## Step:  AIC=1238.04
## MeanAFC ~ FSH + E2 + MaxDailyGn + Oocytes + Embryos
##
##              Df Sum of Sq   RSS    AIC
## - Embryos     1     32.29 13258 1236.8
## <none>                   13226 1238.0
## - FSH         1    132.25 13358 1239.4
## + TotalGn     1     14.55 13212 1239.7
## + Age         1      5.38 13221 1239.9
## + MaxE2       1      4.77 13221 1239.9
## - E2          1    198.35 13424 1241.0
## - Oocytes     1    471.49 13698 1247.7
## - MaxDailyGn  1   1053.50 14280 1261.6
##
## Step:  AIC=1236.85
## MeanAFC ~ FSH + E2 + MaxDailyGn + Oocytes
##
##              Df Sum of Sq   RSS    AIC
## <none>                    13258 1236.8
## + Embryos     1     32.29 13226 1238.0
## - FSH         1    133.22 13392 1238.2
## + TotalGn     1     14.52 13244 1238.5
## + Age         1      7.34 13251 1238.7
## + MaxE2       1      2.75 13256 1238.8
## - E2          1    198.17 13457 1239.8
## - MaxDailyGn  1   1056.80 14315 1260.4
## - Oocytes     1   1452.92 14711 1269.5
```

```
optimal_model<-lm(MeanAFC ~ FSH + E2 + MaxDailyGn + Oocytes, Fertility)
summary(optimal_model)
```

```
##
## Call:
## lm(formula = MeanAFC ~ FSH + E2 + MaxDailyGn + Oocytes, data = Fertility)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.312  -3.413  -0.923   2.595  32.752
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.801458   2.072504   9.072  < 2e-16 ***
## FSH         -0.372004   0.204916  -1.815   0.0704 .
## E2          -0.051418   0.023222  -2.214   0.0275 *
## MaxDailyGn  -0.017469   0.003417  -5.113 5.40e-07 ***
## Oocytes      0.378927   0.063204   5.995 5.36e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.358 on 328 degrees of freedom
## Multiple R-squared:  0.2761, Adjusted R-squared:  0.2673
## F-statistic: 31.28 on 4 and 328 DF,  p-value: < 2.2e-16
```

g) Compare your model from f) to the models from e). Was the stepwise selection method able to find the best subset of variables for any of the criteria? If so, which one(s)?

The stepwise selection method found the best subset of variables for Cp and for adjusted rsq.

h) Decide on the final model you would want to use to predict `MeanAFC` and justify why you choose it. Are the variables for the best model the same as the ones you identified in part b)?

My final model is MeanAFC = B0 + B1(FSH) + B2(E2) + B3(MaxDailyGn) + B4(Oocytes)

I chose this because it was the model predicted by the stepwise, Cp, and adjusted Rˆ2. This is very similar to the variables I identified in part b, but this model also has FSH, which my original model did not.