week_5_assignment

William Foote

2/10/2021

Sort to training and testing

```
set.seed(824)
dt <- sort(sample(nrow(df), nrow(df) * .8, replace = FALSE)) # Sample randomly to training and test
train <- df[dt, ]
test <- df[-dt, ]
```

Exploring 7-predictor model, look at all combinations, and which does regsubsets suggest?

```
df_exh <- regsubsets(Calories ~ . - Id - ActivityDate, data = train, nvmax = 10,
                        method = "exhaustive")

summary(df_exh)$outmat[7, ] # Get row 7 to see the suggested 7-predictor model.
```

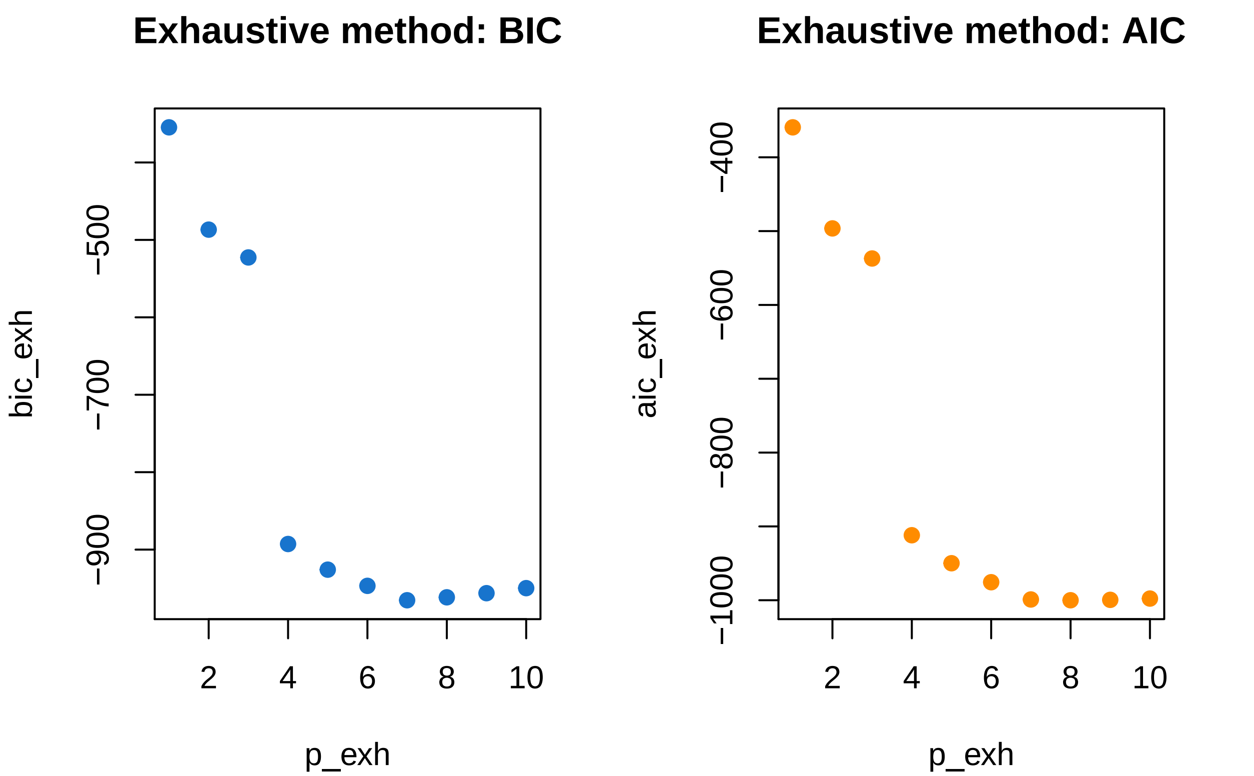
```
##              TotalSteps            TotalDistance LoggedActivitiesDistance
##                     "*"                      "*"                      " "
##      VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                     "*"                      "*"                      " "
##  SedentaryActiveDistance        VeryActiveMinutes      FairlyActiveMinutes
##                     " "                      "*"                      "*"
##    LightlyActiveMinutes         SedentaryMinutes
##                     " "                      "*"
```

The exhaustive method looks at all combinations of all the variables, and can show which models have the highest r-squared for each combination of 1, 2, . . . , 10 variables.

The second line of code shows which 7-predictor model of all the combinations of predictors has the highest R^2.

Is 7-predictor the optimal number, though?

```r
bic_exh <- summary(df_exh)$bic
p_exh <- 1:10
n <- 894 # for 58 counties * 5 years
aic_exh <- bic_exh - log(n) * (p_exh) + 2 * p_exh
aic_exh <- bic_exh - log(n) * (p_exh) + 2 * p_exh
par(mfrow = c(1, 2))
plot(p_exh, bic_exh, col = "dodgerblue3", pch = 19, main = "Exhaustive method: BIC")
plot(p_exh, aic_exh, col = "darkorange", pch = 19, main = "Exhaustive method: AIC")
```




```r
rbind("AIC" = aic_exh, "BIC" = bic_exh)
```

```
##          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## AIC -359.4468 -496.3465 -537.0729 -911.9622 -949.8798 -975.5997 -998.9788
## BIC -354.6511 -486.7551 -522.6858 -892.7794 -925.9013 -946.8254 -965.4089
##           [,8]      [,9]     [,10]
## AIC -1000.0088 -999.4651 -997.7567
## BIC  -961.6432 -956.3037 -949.7997
```

```r
min("AIC" = aic_exh)
```

```
## [1] -1000.009
```

```r
min("BIC" = bic_exh)
```

```
## [1] -965.4089
```

Around 7 variables, both AIC and BIC plateau in terms of continuing their decrease. The returns are diminishing past this point in my opinion (not much increase in R^2 for a much more complex model; tradeoff isn't worth it necessarily).

Making the models for the suggested 7-variable model with regsubsets output.

```
m1 <- lm(Calories ~ TotalSteps + TotalDistance + VeryActiveDistance + ModeratelyActiveDistance + VeryAc
summary(m1)
```

```
##
## Call:
## lm(formula = Calories ~ TotalSteps + TotalDistance + VeryActiveDistance +
##     ModeratelyActiveDistance + VeryActiveMinutes + FairlyActiveMinutes +
##     SedentaryMinutes, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1735.51  -225.61     6.68   241.07  1135.89
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1253.42008   64.42162  19.457  < 2e-16 ***
## TotalSteps                 -0.38584    0.01749 -22.064  < 2e-16 ***
## TotalDistance             689.66887   24.49214  28.159  < 2e-16 ***
## VeryActiveDistance       -308.13332   13.18839 -23.364  < 2e-16 ***
## ModeratelyActiveDistance -329.72866   49.87058  -6.612 7.49e-11 ***
## VeryActiveMinutes          18.09296    0.83377  21.700  < 2e-16 ***
## FairlyActiveMinutes        11.14526    2.21505   5.032 6.18e-07 ***
## SedentaryMinutes            0.33478    0.04979   6.724 3.64e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 357.2 on 707 degrees of freedom
## Multiple R-squared:  0.7592, Adjusted R-squared:  0.7568
## F-statistic: 318.4 on 7 and 707 DF,  p-value: < 2.2e-16
```

```
c("MSE of m1: " = anova(m1)['Residuals', 'Mean Sq'])
```

```
## MSE of m1:
##    127587.2
```

```
m2 <- lm(Calories ~ TotalSteps * TotalDistance * VeryActiveDistance * ModeratelyActiveDistance * VeryAc
# summary(m2)
c("MSE of m2: " = anova(m2)['Residuals', 'Mean Sq'])
```

```
## MSE of m2:
##    93674.81
```

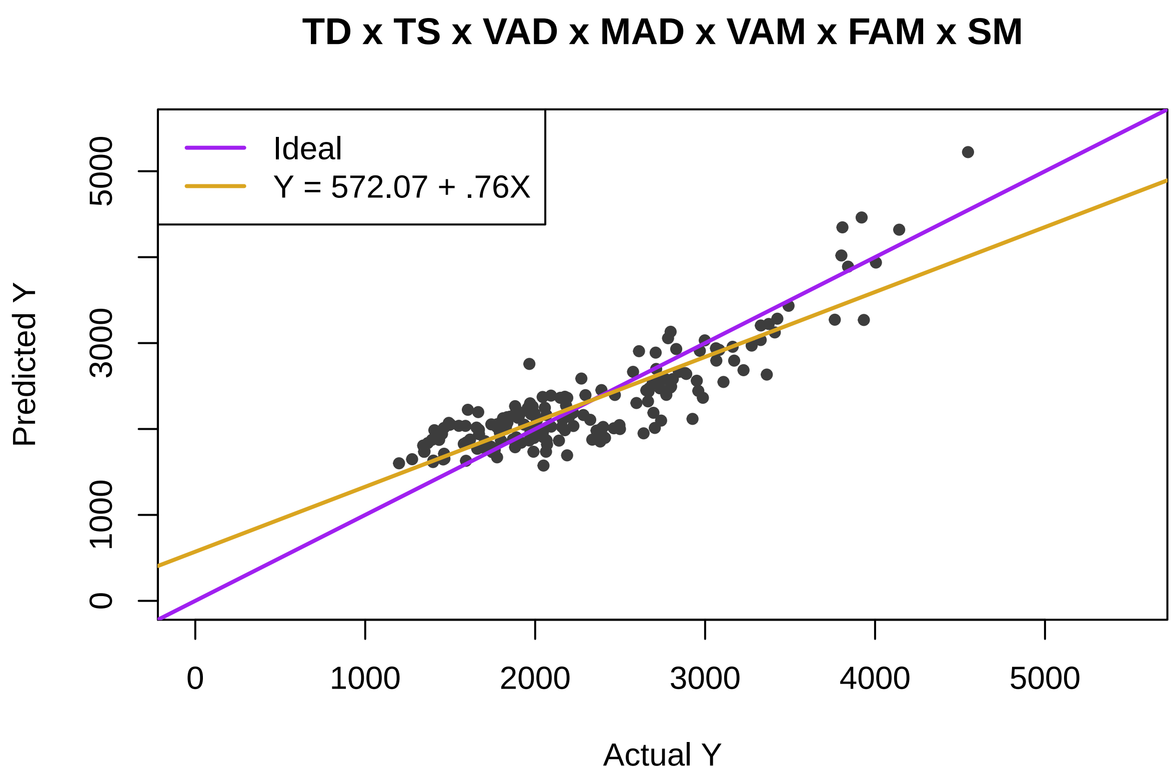
There's a lot that goes on in the summary(m2), so you can remove the # to un-comment it, but in short, the R^2 goes up to 85.32%, and the MSE is 93, 674.81, as outputted. Both values are better, but not sure if it's worth the added terms (of which there are A LOT).

Looking at testing data now

```
pred1 <- predict(m1, newdata = test)
rmse <- sqrt(mean((pred1 - test$Calories)))
c(RMSE = rmse, R2 = summary(m1)$r.squared)
```

```
##      RMSE        R2
## 4.1519385 0.7591927
```

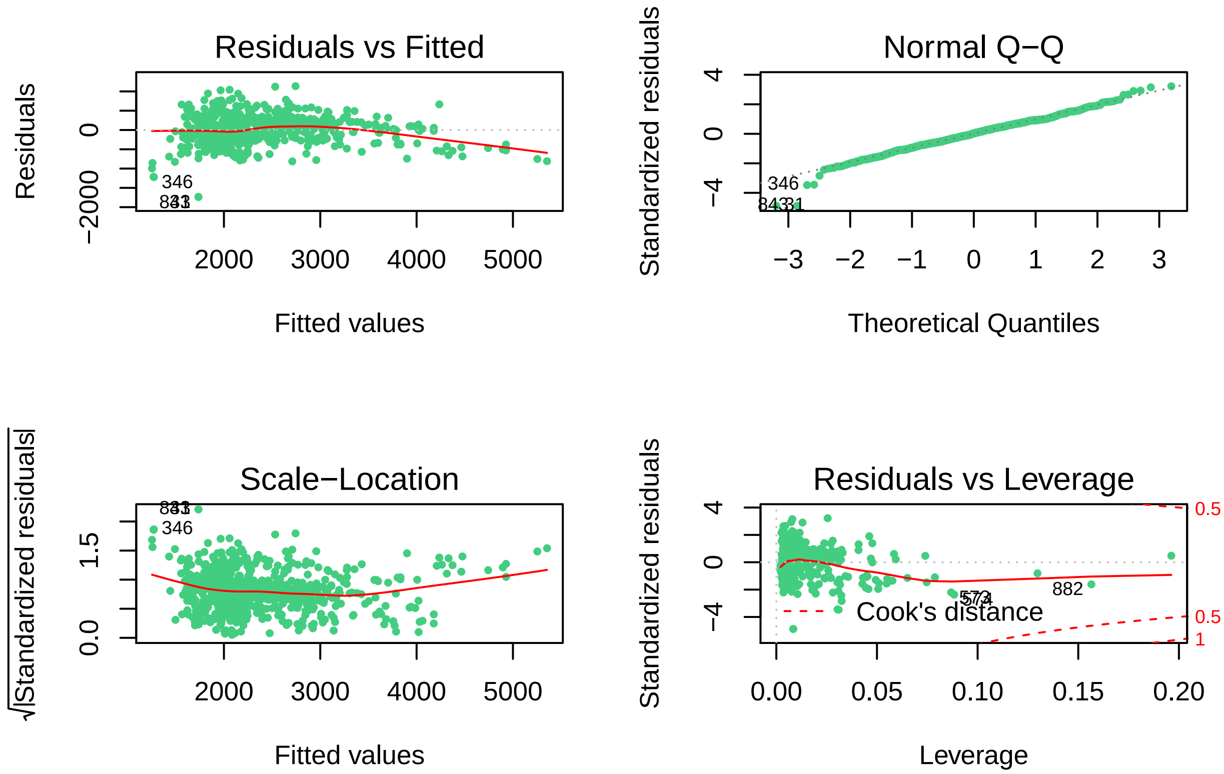
```
m3 <- lm(pred1 ~ test$Calories)

plot(test$Calories, pred1, xlim = c(0, 5500), ylim = c(0, 5500), ylab = "Predicted Y",
     xlab = "Actual Y", main = "TD x TS x VAD x MAD x VAM x FAM x SM",
     pch = 19, cex = .7, col = "grey24")
abline(a = 0, b = 1, col = "purple", lwd = 2)
abline(m3, col = "goldenrod", lwd = 2)
legend("topleft", legend = c("Ideal", "Y = 572.07 + .76X"), col = c("purple", "goldenrod"), lty = 1, lwd
```



Based on RMSE = 4.15, we can conclude that on an average predicted value will be off by 4.15 from the actual value.

Model Validity

```
par(mfrow = c(2, 2))
plot(m1, col = "seagreen3", pch = 19, cex = .50)
```



Linearity: Good, I think. There's a slight decreasing trend at the higher fitted-values, but this could just be because there are less points.

Constant Variation: There isn't a super big strictly decreasing or increasing trend, but there shouldn't be any trend at all so this is potentially worrisome. There also doesn't appear to be a fan shape in the Residuals plot, which would be another form of evidence that the constant variance condition is violated.

Normality: The points follow the Normal-QQ expected line, so this condition is satisfied.

VIF

```r
vif(m1)
```

```
##                TotalSteps             TotalDistance        VeryActiveDistance
##                 42.732604                 50.426702                  7.556582
## ModeratelyActiveDistance          VeryActiveMinutes       FairlyActiveMinutes
##                 11.941367                  4.633705                 11.937188
##          SedentaryMinutes
##                  1.178598
```

```r
head(vif(m2)) # returns all NaNs
```

```
##                TotalSteps             TotalDistance        VeryActiveDistance
##                       NaN                       NaN                       NaN
## ModeratelyActiveDistance          VeryActiveMinutes       FairlyActiveMinutes
##                       NaN                       NaN                       NaN
```

These numbers are all quite high except 2. Values greater than 5 are problematic potentially.

Week 6 Stuff

Looking at VIF vs Model Selection

```
good_model_list <- list(summary(df_exh)$outmat[1, ], summary(df_exh)$outmat[2, ],
     summary(df_exh)$outmat[3, ], summary(df_exh)$outmat[4, ],
     summary(df_exh)$outmat[5, ], summary(df_exh)$outmat[6, ],
     summary(df_exh)$outmat[7, ], summary(df_exh)$outmat[8, ],
     summary(df_exh)$outmat[9, ], summary(df_exh)$outmat[10, ])
good_model_list
```

```
## [[1]]
##                TotalSteps             TotalDistance LoggedActivitiesDistance
##                       " "                       "*"                      " "
##        VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                       " "                       " "                      " "
##   SedentaryActiveDistance         VeryActiveMinutes      FairlyActiveMinutes
##                       " "                       " "                      " "
##      LightlyActiveMinutes          SedentaryMinutes
##                       " "                       " "
##
## [[2]]
##                TotalSteps             TotalDistance LoggedActivitiesDistance
##                       " "                       " "                      " "
##        VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                       " "                       " "                      "*"
##   SedentaryActiveDistance         VeryActiveMinutes      FairlyActiveMinutes
##                       " "                       "*"                      " "
##      LightlyActiveMinutes          SedentaryMinutes
##                       " "                       " "
##
## [[3]]
##                TotalSteps             TotalDistance LoggedActivitiesDistance
##                       " "                       " "                      " "
##        VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                       " "                       " "                      "*"
##   SedentaryActiveDistance         VeryActiveMinutes      FairlyActiveMinutes
##                       " "                       "*"                      " "
##      LightlyActiveMinutes          SedentaryMinutes
##                       "*"                       " "
##
## [[4]]
##                TotalSteps             TotalDistance LoggedActivitiesDistance
##                       "*"                       "*"                      " "
##        VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                       "*"                       " "                      " "
##   SedentaryActiveDistance         VeryActiveMinutes      FairlyActiveMinutes
##                       " "                       "*"                      " "
##      LightlyActiveMinutes          SedentaryMinutes
##                       " "                       " "
##
```

```
## [[5]]
##               TotalSteps            TotalDistance LoggedActivitiesDistance
##                      "*"                      "*"                      " "
##       VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                      "*"                      " "                      " "
##  SedentaryActiveDistance        VeryActiveMinutes      FairlyActiveMinutes
##                      " "                      "*"                      " "
##     LightlyActiveMinutes         SedentaryMinutes
##                      " "                      "*"
##
## [[6]]
##               TotalSteps            TotalDistance LoggedActivitiesDistance
##                      "*"                      "*"                      " "
##       VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                      "*"                      "*"                      " "
##  SedentaryActiveDistance        VeryActiveMinutes      FairlyActiveMinutes
##                      " "                      "*"                      " "
##     LightlyActiveMinutes         SedentaryMinutes
##                      " "                      "*"
##
## [[7]]
##               TotalSteps            TotalDistance LoggedActivitiesDistance
##                      "*"                      "*"                      " "
##       VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                      "*"                      "*"                      " "
##  SedentaryActiveDistance        VeryActiveMinutes      FairlyActiveMinutes
##                      " "                      "*"                      "*"
##     LightlyActiveMinutes         SedentaryMinutes
##                      " "                      "*"
##
## [[8]]
##               TotalSteps            TotalDistance LoggedActivitiesDistance
##                      "*"                      "*"                      " "
##       VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                      "*"                      "*"                      " "
##  SedentaryActiveDistance        VeryActiveMinutes      FairlyActiveMinutes
##                      "*"                      "*"                      "*"
##     LightlyActiveMinutes         SedentaryMinutes
##                      " "                      "*"
##
## [[9]]
##               TotalSteps            TotalDistance LoggedActivitiesDistance
##                      "*"                      "*"                      " "
##       VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                      "*"                      "*"                      " "
##  SedentaryActiveDistance        VeryActiveMinutes      FairlyActiveMinutes
##                      "*"                      "*"                      "*"
##     LightlyActiveMinutes         SedentaryMinutes
##                      "*"                      "*"
##
## [[10]]
##               TotalSteps            TotalDistance LoggedActivitiesDistance
##                      "*"                      "*"                      "*"
##       VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
```

```
##                        "*"                        "*"                        " "
##  SedentaryActiveDistance        VeryActiveMinutes      FairlyActiveMinutes
##                        "*"                        "*"                        "*"
##     LightlyActiveMinutes         SedentaryMinutes
##                        "*"                        "*"
```

```
gml_atomic <- lapply(good_model_list, as.character)
gml_atomic
```

```
## [[1]]
##  [1] " " "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
##
## [[2]]
##  [1] " " " " " " " " " " " " " " " " " " " " "*" " " " " "*" " " " " " " " " " " " "
##
## [[3]]
##  [1] " " " " " " " " " " " " " " " " " " " " "*" " " " " "*" " " " " "*" " " " "
##
## [[4]]
##  [1] "*" "*" " " " " "*" " " " " " " " " " " " " "*" " " " " " " " " " " " "
##
## [[5]]
##  [1] "*" "*" " " " " "*" " " " " " " " " " " " " "*" " " " " " " " " "*"
##
## [[6]]
##  [1] "*" "*" " " " " "*" "*" " " " " " " " " "*" " " " " " " " " "*"
##
## [[7]]
##  [1] "*" "*" " " " " "*" "*" " " " " " " " " "*" "*" " " " " "*"
##
## [[8]]
##  [1] "*" "*" " " " " "*" "*" " " " " "*" "*" "*" " " " " "*"
##
## [[9]]
##  [1] "*" "*" " " " " "*" "*" " " " " "*" "*" "*" "*" "*"
##
## [[10]]
##  [1] "*" "*" "*" "*" "*" " " " " "*" "*" "*" "*" "*"
```

```
gml_desireable_x <- list("m1" = which(gml_atomic[[1]] == "*"))
gml_desireable_x
```

```
## $m1
## [1] 2
```

```
for(i in seq_len(9)) { # Get the rest of the indices of the desired variables for each model
  gml_desireable_x[[paste(c("m", i + 1), collapse = "")]] <- which(gml_atomic[[i + 1]] == "*")
}

for(i in seq_len(10)) { # Convert variable number to column number from df
  gml_desireable_x[[i]] <- gml_desireable_x[[i]] + 2
}
```

```r
gml_desireable_x
```

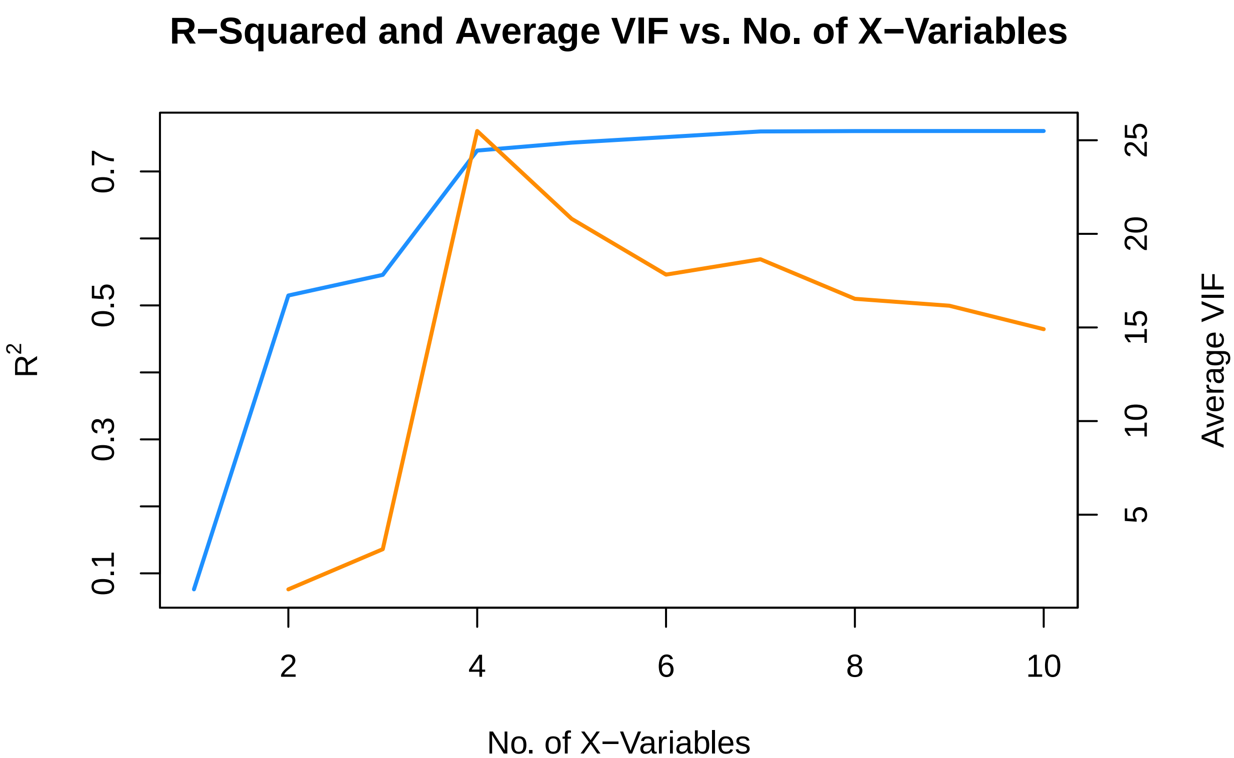
```
## $m1
## [1] 4
##
## $m2
## [1]  8 10
##
## $m3
## [1]  8 10 12
##
## $m4
## [1]  3  4  6 10
##
## $m5
## [1]  3  4  6 10 13
##
## $m6
## [1]  3  4  6  7 10 13
##
## $m7
## [1]  3  4  6  7 10 11 13
##
## $m8
## [1]  3  4  6  7  9 10 11 13
##
## $m9
## [1]  3  4  6  7  9 10 11 12 13
##
## $m10
##  [1]  3  4  5  6  7  9 10 11 12 13
```

```r
gml_models <- data.frame("r.squared" = summary(lm(df[, 14] ~ df[, 2]))$r.squared, "avg_vif" = NA) # df[

for(i in seq_len(9)) {
  model <- lm(Calories ~ ., data = cbind("Calories" = df[, 14], df[, gml_desireable_x[[i + 1]]]))
  gml_models[i + 1, 1] <- summary(model)$r.squared
  gml_models[i + 1, 2] <- mean(vif(model))
}

gml_models
```

```
##    r.squared   avg_vif
## 1  0.0760970        NA
## 2  0.5147998  1.012150
## 3  0.5456496  3.157497
## 4  0.7311896 25.493296
## 5  0.7430198 20.804148
## 6  0.7512721 17.822342
## 7  0.7597161 18.643556
## 8  0.7602858 16.530512
## 9  0.7603637 16.161532
## 10 0.7603948 14.908425
```

Look at VIF and R-Squared Changes As More Variables are Added

```
par(mar = c(5, 4, 4, 4) + 0.3)  # Leave space for avg_vif axis
plot(gml_models$r.squared, type = "l", col = "dodgerblue", ylab = expression(R^2), xlab = "No. of X-Vari
par(new = TRUE)
plot(gml_models$avg_vif, type = "l", col = "darkorange", axes = FALSE, bty = "n", xlab = "", ylab = "",
     lwd = 2)
mtext("Average VIF", side = 4, line = 3)
axis(side = 4, at = pretty(range(gml_models$avg_vif, na.rm = TRUE)))
```



The three-predictor model looks best, as average VIF skyrockets in any model with more than 3 variables. R-Squared is about 20% lower, but worth the tradeoff in my opinion.

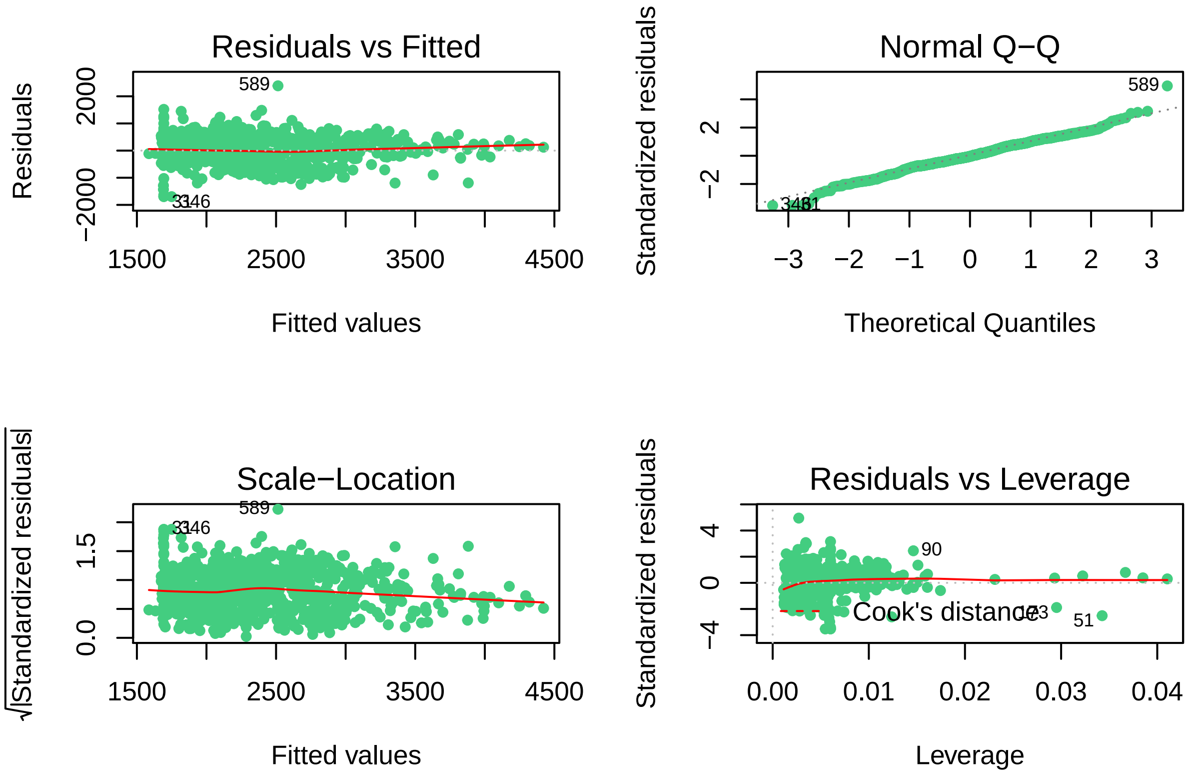
Investigate this model

```
m3_opt <- lm(Calories ~ ., data = cbind("Calories" = df[, 14], df[, gml_desireable_x[[3]]]))
vif(m3_opt)
```

```
##  LightActiveDistance    VeryActiveMinutes LightlyActiveMinutes
##             4.233047             1.056984             4.182460
```

Diagnostics

```r
par(mfrow = c(2, 2))
plot(m3_opt, cex = .75, pch = 19, col = "seagreen3")
```



All of the diagnostic plots look pretty good for this model as well.