Rough Draft Q2

Considering the constraint of the Climate Change Act and other government targets, produce a model forecasting the carbon emissions of the evolving profile of vehicles on the UK road network.

```r
# load data
all <- read.csv("GB_all_data.csv")
colnames(all)[1] <- "Year"

# take the date out - this new df will be used for subset regression coming up
df <- all[,-c(1, 2, 3)]

# remove 2019
df <- df[1:25,]

# turn into df
df <- data.frame(df)
```

```r
# use all predictors
m1 <- lm(Total_GHGs ~., data = df)
```

1.) Best subset regression selection

The number of models that this procedure fits multiplies quickly. If you have 10 independent variables, it fits 1024 models. However, if you have 20 variables, it fits 1,048,576 models

```r
# https://statisticsbyjim.com/regression/guide-stepwise-best-subsets-regression/
# https://olsrr.rsquaredacademy.com/articles/variable_selection.html
# change the ols() function, there are cool ones to pick from
ols_step_best_subset(m1)
```

```
##                                                                      Best Subsets Regressio
## ---------------------------------------------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------------------------------------------
##      1         Total_Production
##      2         Unemployment_rate Total_Production
##      3         Hybrid_Electric Range_Extended_Electric Total_Vehicles
##      4         Diesel Hybrid_Electric Range_Extended_Electric Total_Production
##      5         Hybrid_Electric BE Gas Other Total_Vehicles
##      6         Diesel Hybrid_Electric Plug.in_HE Gas Other Unemployment_rate
##      7         Diesel Hybrid_Electric Plug.in_HE Fuel_Cell_Electric Gas Other Unemployment_rate
```

```
##      8         Petrol Hybrid_Electric Plug.in_HE Fuel_Cell_Electric Gas Other Total_Vehicles Unemplo
##      9         Petrol Diesel Hybrid_Electric Fuel_Cell_Electric Gas Other Total_Vehicles Unemploymen
##     10         Petrol Diesel Hybrid_Electric Plug.in_HE BE Fuel_Cell_Electric Other Total_Vehicles U
##     11         Petrol Diesel Hybrid_Electric Plug.in_HE BE Range_Extended_Electric Gas Total_Vehicle
##     12         Petrol Diesel Hybrid_Electric Plug.in_HE BE Range_Extended_Electric Fuel_Cell_Electri
##     13         Petrol Diesel Hybrid_Electric Plug.in_HE BE Range_Extended_Electric Fuel_Cell_Electri
##     14         Petrol Diesel Hybrid_Electric Plug.in_HE BE Range_Extended_Electric Fuel_Cell_Electri
## ----------------------------------------------------------------------------------------------------
##
##                                                           Subsets Regression Summary
## ----------------------------------------------------------------------------------------------------
##                       Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)        AIC        SBIC       SBC           MSEP
## ----------------------------------------------------------------------------------------------------
##   1        0.7611      0.7507      0.7122    96.2483    454.5695      NA     458.2262     98685865.023
##   2        0.8019      0.7839       0.745    78.2418    451.8922      NA     456.7677     85744141.566
##   3        0.9019      0.8879      0.8712    31.1448    436.3175      NA     442.4119     44574854.511
##   4        0.9313      0.9176      0.8828    18.7141    429.4101      NA     436.7234     32857117.221
##   5        0.9612      0.9510      0.9277     6.0237    417.1043      NA     425.6364     19570107.887
##   6        0.9686      0.9581      0.9322     4.4187    413.8517      NA     423.6027     16794590.317
##   7        0.9716      0.9598      0.9419     4.9592    413.3656      NA     424.3355     16155162.644
##   8        0.9744      0.9615      0.9339     5.5875    412.7797      NA     424.9685     15538834.975
##   9        0.9760      0.9616      0.9246     6.7696    413.1002      NA     426.5078     15566981.426
##  10        0.9765      0.9598      0.9188     8.5204    414.5652      NA     429.1917     16409538.845
##  11        0.9773      0.9581       0.922    10.1459    415.7390      NA     431.5843     17199064.174
##  12        0.9775      0.9549      0.9091    12.0551    417.5345      NA     434.5988     18609783.017
##  13        0.9776      0.9511       0.889    14.0000    419.4096      NA     437.6927     20368712.815
##  14        0.9776      0.9511       0.889    14.0000    421.4096      NA     440.9116     20368712.815
## ----------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

This takes a very long time if you use all 14 X variables.

```r
# based on OLS results we selected these two variables
m2 <- lm(Total_GHGs ~ Hybrid_Electric + Range_Extended_Electric + Total_Vehicles, data = df)

# check out model
summary(m2)
```

```
##
## Call:
## lm(formula = Total_GHGs ~ Hybrid_Electric + Range_Extended_Electric +
##     Total_Vehicles, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2279.4  -766.1   -55.6   638.3  3646.0
```

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             8.856e+04  4.103e+03  21.585 8.12e-16 ***
## Hybrid_Electric        -1.309e-01  1.004e-02 -13.042 1.55e-11 ***
## Range_Extended_Electric  4.263e+00  3.599e-01  11.845 9.25e-11 ***
## Total_Vehicles          1.203e-03  1.662e-04   7.241 3.92e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1332 on 21 degrees of freedom
## Multiple R-squared:  0.9019, Adjusted R-squared:  0.8879
## F-statistic: 64.36 on 3 and 21 DF,  p-value: 9.334e-11
```

```r
# https://www.researchgate.net/post/How_high_of_VIF_value_in_regression_can_be_accepted
# check for colinearity
vif(m2)
```

```
##         Hybrid_Electric Range_Extended_Electric          Total_Vehicles
##               18.719487               11.993388                3.521579
```

A VIF above 5 is not good, a VIF above 10 is very bad, so we will go to the 2 X-variable model, which is Unemployment__rate Total__Production = Yhat.

```r
# based on OLS results we selected these two variables
m3 <- lm(Total_GHGs ~ Unemployment_rate + Total_Production, data = df)

# check out model
summary(m3)
```

```
##
## Call:
## lm(formula = Total_GHGs ~ Unemployment_rate + Total_Production,
##     data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3749  -1550    289   1284   2982
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        34168.7    11862.8   2.880  0.00869 **
## Unemployment_rate   -600.3      282.1  -2.128  0.04483 *
## Total_Production     842.0      110.0   7.655 1.22e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1850 on 22 degrees of freedom
## Multiple R-squared:  0.8019, Adjusted R-squared:  0.7839
## F-statistic: 44.52 on 2 and 22 DF,  p-value: 1.847e-08
```

```r
# https://www.researchgate.net/post/How_high_of_VIF_value_in_regression_can_be_accepted
# check for colinearity
vif(m3)
```

```
## Unemployment_rate  Total_Production
##          1.175879          1.175879
```

2.) Forecasting our X variables

https://otexts.com/fpp2/the-forecast-package-in-r.html

If the first argument is of class ts, it returns forecasts from the automatic ETS algorithm discussed in Chapter 7.

```r
# use only needed data
future_df <- all[,c("Year", "Unemployment_rate", "Total_Production")]

# turn into time series data
future_df <- ts(future_df)

# forecast 4 ahead (4 is arbitrary)
fore_df <- forecast(future_df, h = 4) %>% data.frame()

# keep only point forecast
fore_df <- fore_df[,1:3]

# turn df into long shape
fore_df_long <- pivot_wider(fore_df, names_from = Series, values_from = Point.Forecast)
```

```r
coef(m3)
```

```
##       (Intercept) Unemployment_rate  Total_Production
##        34168.6941         -600.2673          841.9804
```

Now add those to df as constants.

```r
fore_df_long$Intercept <- 34168.6941
fore_df_long$UR_coef <- -600.2673
fore_df_long$TP_coef <- 841.9804
```

3.) Final predictions

```r
# predictions
final <- fore_df_long %>%
  mutate(Y_hat = (Intercept + (Unemployment_rate * UR_coef) + (Total_Production  * TP_coef)))


head(final)
```

```
## # A tibble: 4 x 8
##   Time   Year Unemployment_ra~ Total_Production Intercept UR_coef TP_coef  Y_hat
##   <chr> <dbl>            <dbl>            <dbl>     <dbl>   <dbl>   <dbl>  <dbl>
## 1 27     2020             3.48             97.9    34169.   -600.    842. 1.15e5
## 2 28     2021             3.17             97.9    34169.   -600.    842. 1.15e5
## 3 29     2022             2.85             97.9    34169.   -600.    842. 1.15e5
## 4 30     2023             2.54             97.9    34169.   -600.    842. 1.15e5
```

```r
# just see the predicted values
print(final$Y_hat)
```

```
## [1] 114507.3 114696.2 114885.1 115073.9
```

From 2015 - 2018 our data was * 111,973.846 * 114,585.172 * 114,785.965 * 113,280.299

Thoughts

- To be honest, not a big fan of using total production for an X variable. Maybe a lag variable instead?
- Changing the forecasting approach. This was just a quick example
- X columns with a lot of 0's? not sure it's good to have, if okay to have we can add other ones such as Electric charging points