inlämning 3

Umut Arslan

2023-03-06

Datasett taget ifrån bibloteket ISLR2, Auto.

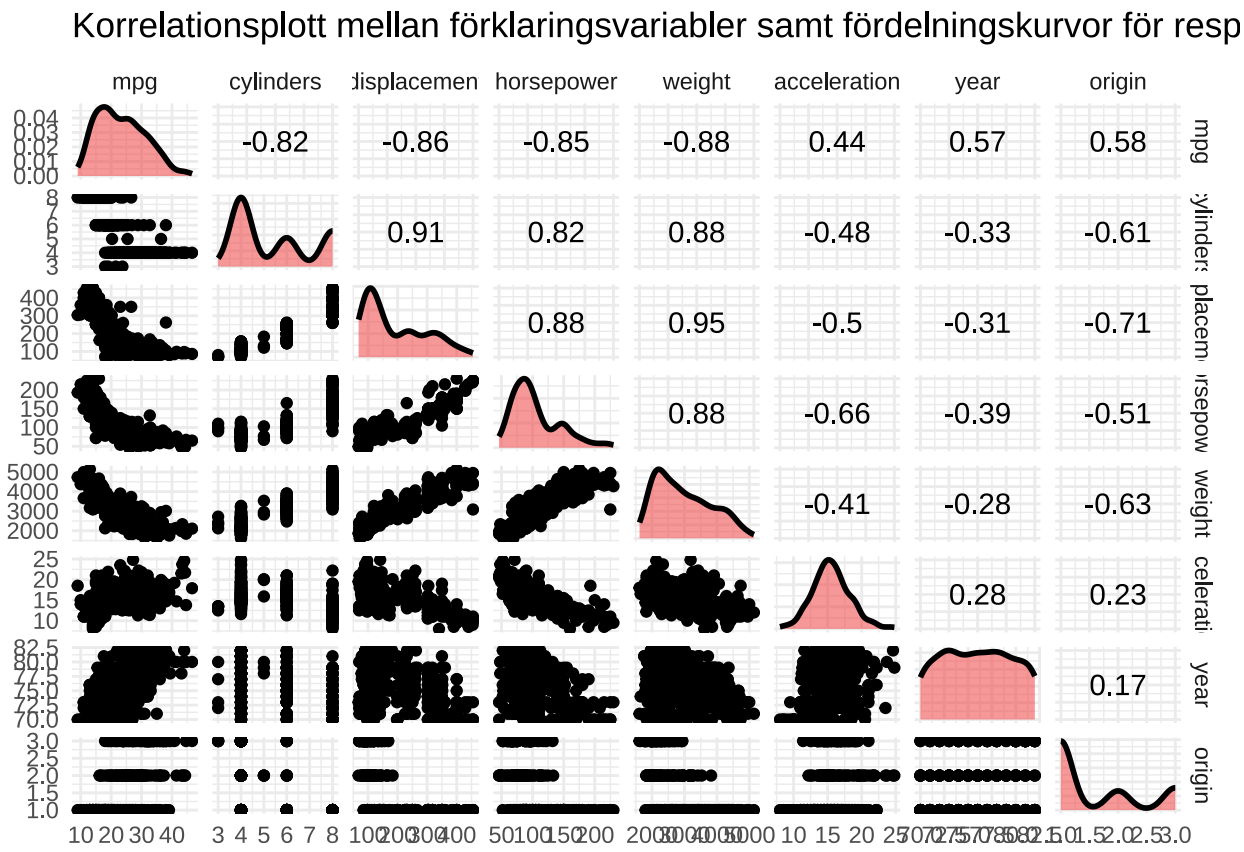
Auto innehåller 9 olika kolumner där mpg, displacement och acceleration är av datatypen Numerical, kontinuerliga siffror.

Cylinders, horsepower, weight, year och origin är av datatypen Integer, heltal.

Name är av datatypen Factor

Mpg är hur många "miles" per "gallon" bilen drar, cylinders är antalet cylindrar bilen har, displacment är slagvolym för bilen, horsepower är hästkraften för bilen, weight är vikten på bilen, acceleration är definierad som tiden att accelerera från 0 till ungefär 96km/h, year är årsmodellen på bilen, origin är vart bilen kommer ifrån (1. Amerika, 2. Europa, 3. Japan), name är bilens namn.

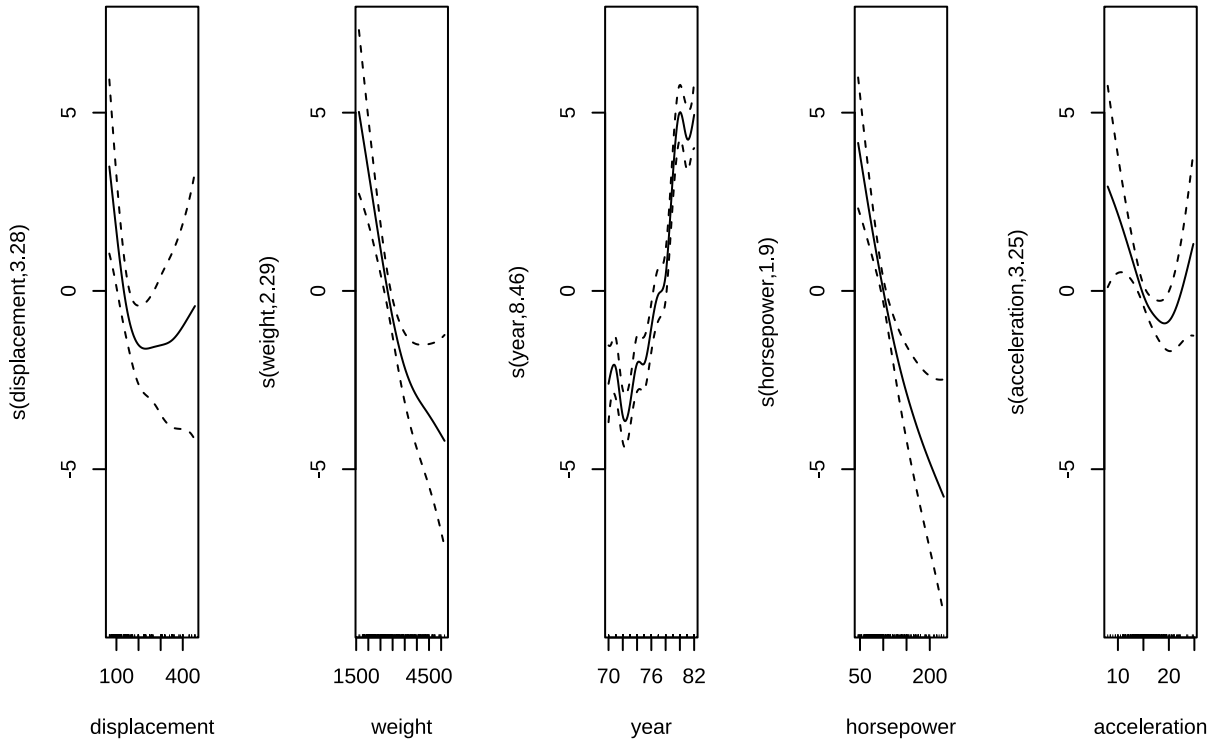
Nedan kan vi se att cylinders och origin ser ut att vara kategoriska variabler. När vi gör våran gam-modell kommer vi att sätta dessa som "dummie" variabler.



GAM av datasett Auto

När vi undersöker diverse modeller så kommer jag fram till att man skulle kunna excludera förklaringsvariabeln origin beroende på vilken gräns man har, jämför man modellen med alla variabler mot modell med alla minus origin så blir resultatet försummbart.

Nedan kan vi se att acceleration samt displacement har ett icke-linjärt samband medan resterande av förklaringsvariablerna ser ut att ha ett linjärt samband med mpg. Modellen visar även att antalet cylindrar påverkar mpg negativt, alltså desto fler cylindrar tenderar att minska konsumptionen av bränsle. Samma sak gäller för weight och horsepower. Year ger ökad mpg.



Titanic datasett

Datasettet innehåller Passangerid, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin och Embarked

- PassangerId är passangerarnummret
- Survivad (0 = No, 1 = Yes)
- Pclass - Vilken klass passageraren hade, där 1 är första klass, 2 är andra och tre är tredje klass.
- Name är personernas namn
- Sex är vilket kön personen hade
- Age är vilken ålder personen hade
- SibSp är antalet syskon/sambo personen hade på båten
- Parch är antalet föräldrar/barn personen hade på båten
- Ticket är vilket nummer personen hade på biljetten
- Fare är avgiften personen betalade för biljetten

- Cabin är vart personen hade sin hytt
- Embarked är vilket hamn passageraren steg in på båten (C = Cherbourg; Q = Queenstown; S = Southampton)

Jag kommer att transformera Name och Cabin till title och Decck. title kommer att innehålla Mr, Miss, Mrs eller Master och Deck kommer innehålla bokstaven A till G beroende på vart dom hade sin hytt.

Name och Cabin kommer att tas bort från datasettet.

Vi börjar med att undersöka om det saknas datapunkter.

```
## # A tibble: 12 x 17
##    skim_type skim_vari~1 n_mis~2 compl~3 chara~4 chara~5 chara~6 chara~7 chara~8
##    <chr>     <chr>         <int>   <dbl>   <int>   <int>   <int>   <int>   <int>
##  1 character Sex               0   1           4       6       0       2       0
##  2 character Ticket            0   1           3      18       0     681       0
##  3 character Embarked          0   1           0       1       2       4       0
##  4 character title             0   1           2       7       0       4       0
##  5 character Deck              0   1           1       7       0       8       0
##  6 numeric   PassengerId       0   1          NA      NA      NA      NA      NA
##  7 numeric   Survived          0   1          NA      NA      NA      NA      NA
##  8 numeric   Pclass            0   1          NA      NA      NA      NA      NA
##  9 numeric   Age             177   0.801      NA      NA      NA      NA      NA
## 10 numeric   SibSp             0   1          NA      NA      NA      NA      NA
## 11 numeric   Parch             0   1          NA      NA      NA      NA      NA
## 12 numeric   Fare              0   1          NA      NA      NA      NA      NA
## # ... with 8 more variables: numeric.mean <dbl>, numeric.sd <dbl>,
## #   numeric.p0 <dbl>, numeric.p25 <dbl>, numeric.p50 <dbl>, numeric.p75 <dbl>,
## #   numeric.p100 <dbl>, numeric.hist <chr>, and abbreviated variable names
## #   1: skim_variable, 2: n_missing, 3: complete_rate, 4: character.min,
## #   5: character.max, 6: character.empty, 7: character.n_unique,
## #   8: character.whitespace
```

Figuren ovan visar oss att cirka 20% av variabeln “Age” saknas i datat.

Antingen tar vi bort dom raderna, det vill säga 177 rader, eller så kan vi “impute missing data”. Vi testar att göra det sistnämda.

Nedan ser vi att det inte saknas några missing values från datasettet längre, detta gjordes med hjälp av step_impute_knn, där jag valde de 10 närmaste grannarna, dvs k = 10.

```
## Warning: There was 1 warning in `dplyr::summarize()`.
## i In argument: `dplyr::across(tidyselect::any_of(variable_names),
##   mangled_skimmers$funs)`.
## i In group 0: .
## Caused by warning:
## ! There was 1 warning in `dplyr::summarize()`.
## i In argument: `dplyr::across(tidyselect::any_of(variable_names),
##   mangled_skimmers$funs)`.
## Caused by warning in `sorted_count()`:
## ! Variable contains value(s) of "" that have been converted to "empty".
```

```
## # A tibble: 12 x 15
##    skim_type skim_vari~1 n_mis~2 compl~3 facto~4 facto~5 facto~6 numer~7 numer~8
##    <chr>     <chr>         <int>   <dbl> <lgl>     <int> <chr>     <dbl>   <dbl>
```

```
##  1 factor    Pclass            0       1 FALSE         3 3: 491~  NA      NA
##  2 factor    Sex               0       1 FALSE         2 mal: 5~  NA      NA
##  3 factor    Ticket            0       1 FALSE       681 160: 7~  NA      NA
##  4 factor    Embarked          0       1 FALSE         4 S: 644~  NA      NA
##  5 factor    title             0       1 FALSE         4 Mr: 64~  NA      NA
##  6 factor    Deck              0       1 FALSE         8 Unk: 6~  NA      NA
##  7 factor    Survived          0       1 FALSE         2 0: 549~  NA      NA
##  8 numeric   PassengerId       0       1 NA           NA <NA>    446     257.
##  9 numeric   Age               0       1 NA           NA <NA>     29.7    13.6
## 10 numeric   SibSp             0       1 NA           NA <NA>      0.523   1.10
## 11 numeric   Parch             0       1 NA           NA <NA>      0.382   0.806
## 12 numeric   Fare              0       1 NA           NA <NA>     32.2    49.7
## # ... with 6 more variables: numeric.p0 <dbl>, numeric.p25 <dbl>,
## #   numeric.p50 <dbl>, numeric.p75 <dbl>, numeric.p100 <dbl>,
## #   numeric.hist <chr>, and abbreviated variable names 1: skim_variable,
## #   2: n_missing, 3: complete_rate, 4: factor.ordered, 5: factor.n_unique,
## #   6: factor.top_counts, 7: numeric.mean, 8: numeric.sd
```

##Random Forest importance

Baserat på utskriften nedan ser vi att den förklaringsvariabel som förklarar responsvariabeln bäst är Sex. Sedan kommer Ticket, Age, Fear, PassangerID, title, Pclass, Deck, SibSp, Embarked och sist Parch.

```
##             MeanDecreaseGini
## PassengerId        44.979283
## Pclass             25.135059
## Sex                87.876190
## Age                47.987880
## SibSp              16.139747
## Parch               8.967873
## Ticket             57.152000
## Fare               52.610225
## Embarked            8.712159
## title              27.475295
## Deck               20.189963
```