

Recommended Exercises (Module 2)

Jim Totland

1/16/2021

[Link to problem set](#)

Problem 1

- a) Weather forecasting. Response: “Sunny”, “Cloudy”, “Rain” etc. Predictors: Air pressure, temperature, and the weather of the previous day(s). The goal is to predict.
- b) Battery life of a phone. Response: Time until the phone is dead. Predictors: Screen size, Battery specs, Processor etc. Both prediction and inference are relevant here. Given a phone, we want to be able to predict what the battery life will be, based to the predictors, but from the regression we will also be able to infer which predictors are most significant.

Problem 2

- a) In this example, the more flexible methods have a smaller test MSE. But at some point the test MSE start to increase monotonically with the flexibility. This is a result of overfitting.
- b) The variance refers to how much \hat{f} would change if we used another set of training data. A small variance could indicate that a rigid method has been used, implying that the data is most likely underfitted.
- c) Bias generally decreases with flexibility, which indicates that a very low bias is connected to overfitting the data.

Problem 3

```
library(ISLR)
data(Auto)
```

- a) Use the `glimpse` function from the tidyverse:

```
glimpse(Auto)
```

```
## Rows: 392
## Columns: 9
## $ mpg      <dbl> 18, 15, 18, 16, 17, 15, 14, 14, 14, 15, 15, 14, 15, 14...
## $ cylinders <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 4, 6, 6, 6, ...
## $ displacement <dbl> 307, 350, 318, 304, 302, 429, 454, 440, 455, 390, 383,...
## $ horsepower <dbl> 130, 165, 150, 150, 140, 198, 220, 215, 225, 190, 170,...
## $ weight      <dbl> 3504, 3693, 3436, 3433, 3449, 4341, 4354, 4312, 4425, ...
## $ acceleration <dbl> 12.0, 11.5, 11.0, 12.0, 10.5, 10.0, 9.0, 8.5, 10.0, 8....
## $ year        <dbl> 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70...
## $ origin      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, ...
## $ name        <fct> chevrolet chevelle malibu, buick skylark 320, plymouth...
```

The data has dimensions 392×9 . All predictors except `name` are quantitative, although some of them may also be treated as categorical.

b) The range is found by applying the `range()` function. For example:

```
range(Auto$mpg)
```

```
## [1] 9.0 46.6
```

c) The mean and standard deviation can be found in the following way:

```
for (i in 1:8) {
  print(summarise(Auto, mean = mean(Auto[,i]), sd = sd(Auto[,i])))
}
```

```
##      mean      sd
## 1 23.44592 7.805007
##      mean      sd
## 1 5.471939 1.705783
##      mean      sd
## 1 194.412 104.644
##      mean      sd
## 1 104.4694 38.49116
##      mean      sd
## 1 2977.584 849.4026
##      mean      sd
## 1 15.54133 2.758864
##      mean      sd
## 1 75.97959 3.683737
##      mean      sd
## 1 1.576531 0.8055182
```

d) Possible, though not very clean, solution:

```
ReducedAuto <- Auto[-(10:85),]

for (i in 1:8) {
  print(summarise(ReducedAuto, mean = mean(ReducedAuto[,i]),
                  sd = sd(ReducedAuto[,i]),
                  range = range(ReducedAuto[,i])))
}
```

```
##      mean      sd range
## 1 24.40443 7.867283 11.0
## 2 24.40443 7.867283 46.6
##      mean      sd range
## 1 5.373418 1.654179 3
## 2 5.373418 1.654179 8
##      mean      sd range
## 1 187.2405 99.67837 68
## 2 187.2405 99.67837 455
##      mean      sd range
## 1 100.7215 35.70885 46
## 2 100.7215 35.70885 230
##      mean      sd range
## 1 2935.972 811.3002 1649
## 2 2935.972 811.3002 4997
##      mean      sd range
```

```
## 1 15.7269 2.693721 8.5
## 2 15.7269 2.693721 24.8
##      mean      sd range
## 1 77.14557 3.106217 70
## 2 77.14557 3.106217 82
##      mean      sd range
## 1 1.601266 0.81991 1
## 2 1.601266 0.81991 3
```

e)

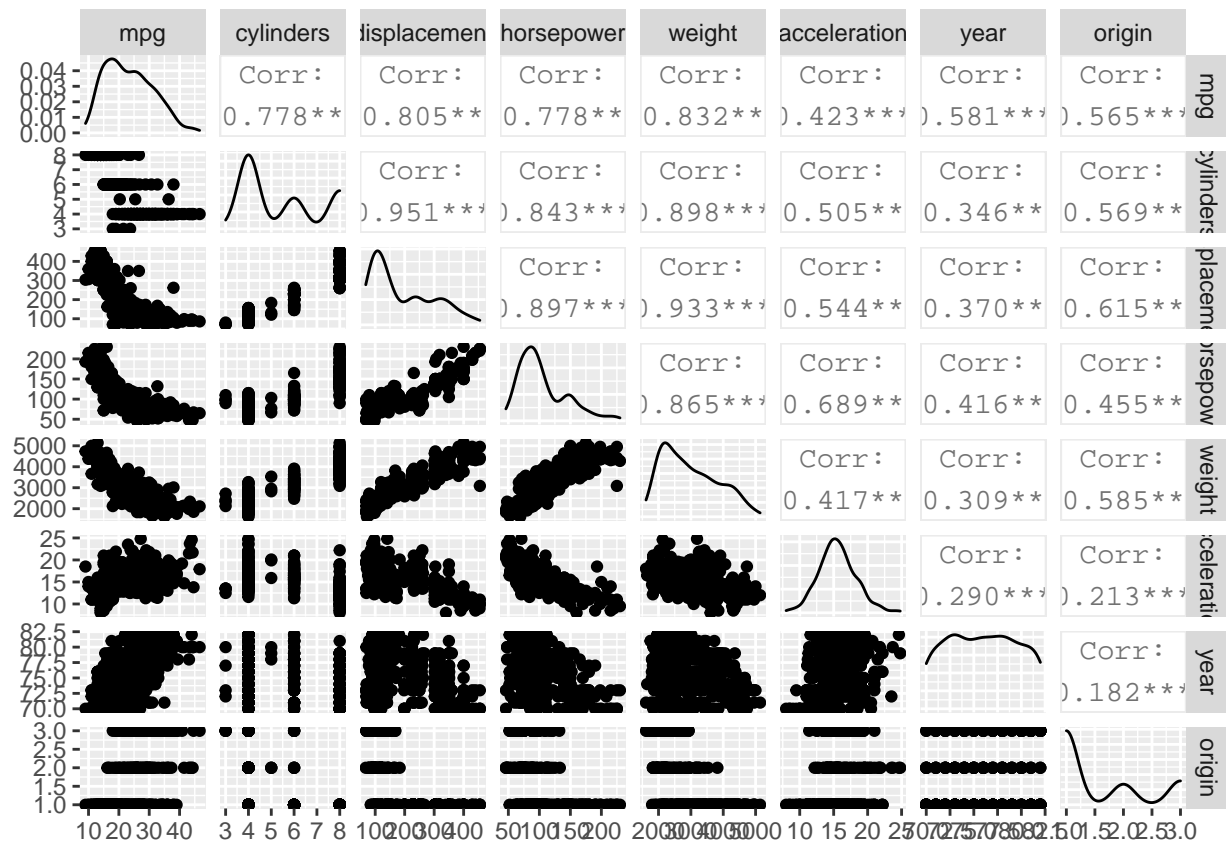
```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
```

```
## method from
```

```
## +.gg ggplot2
```

```
ggpairs(Auto[-9])
```



From the plot we can see that there seems to be a linear relationship between multiple predictors. E.g. weight and displacement have a clearly positive linear relationship. There also seems to be some non-linear relationships, e.g. between mpg and horsepower.