

# WDS, HW 3

Aaron      Neil      Ben      Luke      Tanvi

Due: 10:00PM, July 21, 2021

## Contents

<b>Case study: Automobiles efficiency</b>	<b>2</b>
<b>EDA</b>	<b>2</b>
What effect does <b>time</b> have on <b>MPG</b> ? . . . . .	2
Bring <b>origin</b> into the model . . . . .	3

## Case study: Automobiles efficiency

Are cars being built more efficient? Are Asian cars more efficient than cars built in America or Europe? To answer the questions we will use the `Auto` dataset from ISLR. The original dataset contains 408 observations about cars. It is similar to the `CARS` dataset that we use in our lectures. But it also collects information by years. To get the data, first install the package ISLR. The `Auto` dataset should be loaded automatically. The original data source is here: <https://archive.ics.uci.edu/ml/datasets/auto+mpg>

Get familiar with this dataset first. A good data set should be well documented. Use the command `?ISLR::Auto` to view a description of the dataset. Please add the variable list with names, brief descriptions and units of the variables below.

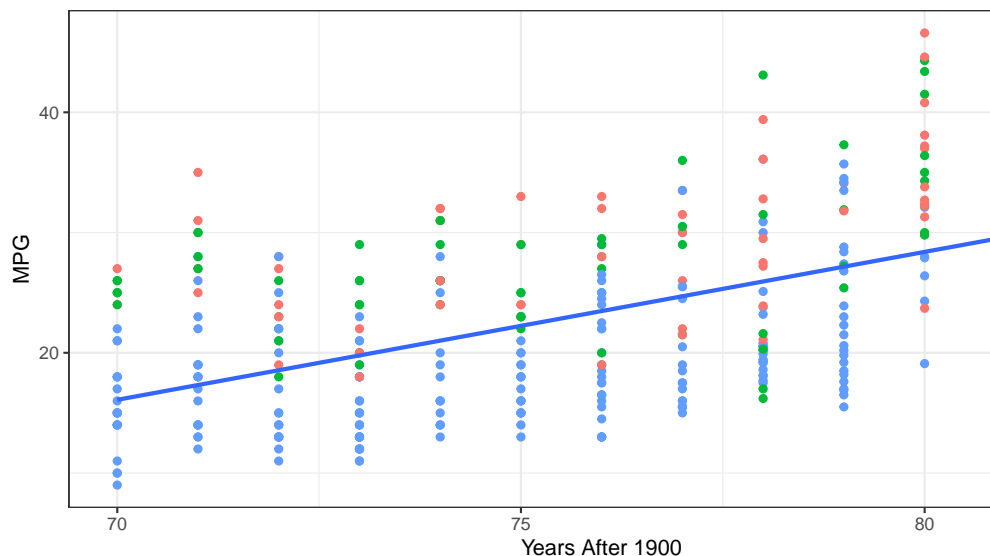
## EDA

Explore the data first.

- i. What is the range of `year`? Why is this important to know?
  - The range of `year` is 12. This is important because it shows the time span of the dataset.
- ii. Should `origin` be a continuous variable? Why or why not. In any case make `origin` a categorical variable.
  - By first looking at `origin`, we may think it is a continuous variable as it is made up of the numbers 1, 2, and 3. However it is a categorical variable and `origin` 1, 2, 3 suggest different origins.
- iii. Do you see any peculiarity in the data?
  - Most of the dataset looks fine, except that the `origin` is named 1, 2, 3, while we would think they would have distinguished names. This makes it rather peculiar.

## What effect does time have on MPG?

- i. Show a scatter plot of `mpg` vs. `year` with the LS line imposed. Does the plot show a positive trend?



- Yes, there is a positive trend.

ii. Now run a simple regression of `mpg` vs. `year` and report R's `summary` output. Is `year` a significant variable at the .05 level? State what effect `year` has on `mpg`, if any, according to this model.

- As `year` has three stars, we can conclude that it as a variable has 0.01 significance. According to this model, when `year` is increased by one, `mpg` increases by 1.2300.

iii. Add `horsepower` on top of the variable `year` to your linear model. Is `year` still a significant variable at the .05 level? Give a precise interpretation of the `year`'s effect found here.

- In this model, when `horsepower` is kept constant, when `year` changes by 1 or increases by 1, the `mpg` increase by 0.65727. `year` still has significance of 0.001.

iv. The two 95% CI's for the coefficient of `year` differ among ii. and iii. How would you explain the difference to a non-statistician?

The two confidence intervals for the coefficient of `year` differ among ii. and iii. because in iii. the linear regression model has to account for both `year` and `horsepower` (keeping `horsepower` constant but still as a factor), whereas in ii. the linear regression model only has to account for the `year`. In turn, the confidence interval for ii. will accept more values into the interval based upon its coefficient, whereas the confidence interval for iii. will accept less values into the interval at a lower coefficient.

v. Create a model with interaction by fitting `lm(mpg ~ year * horsepower)`. Is the interaction effect significant at .05 level? Explain the `year` effect (if any).

- Yes, the interaction effect is significant at .05 level (3 stars). The `year` effect is best expressed as follows.

$$\Delta mpg = (\beta_1 + \beta_3 hp) \cdot \Delta year$$

## Bring origin into the model

Do `mpg`'s differ on average among different `origin`? Fit a linear model with `mpg` vs. `origin`. Report the output.

```
##
## Call:
## lm(formula = mpg ~ data1$origin, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.45  -5.03  -1.03   3.65  18.97
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30.451     0.720   42.31  <2e-16 ***
## data1$origin2  -2.848     1.058   -2.69  0.0074 **
## data1$origin1 -10.417     0.828  -12.59  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.4 on 389 degrees of freedom
## Multiple R-squared:  0.332, Adjusted R-squared:  0.328
## F-statistic: 96.6 on 2 and 389 DF, p-value: <2e-16
```

i. Are `mpg`'s on average different among three regions? Perform a test at .01 level. When you reject the null hypothesis, what have you proved?

- The f statistic produced by `Anova` function is 96.6 which is significantly larger than 1. Since an f statistic of 1 shows the null hypothesis, the very large f statistic shows that we can reject the null hypothesis and say that average of the three `origins` are different in a statistically significant way.

```
## Anova Table (Type II tests)
##
## Response: mpg
##           Sum Sq  Df F value Pr(>F)
## data1$origin   7904   2    96.6 <2e-16 ***
## Residuals    15915 389
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i. Describe on average which `origin` has the highest `mpg` and what it is. Which `origin` has the smallest `mpg` on average and what is it?

- On average, `origin 3` has the highest `mpg` of 30.451, and `origin 1` has the lowest average `mpg` of 20.033.