STATS 2107
Statistical Modelling and Inference II

Practical 6: ANCOVA

Sharon Lee, Matt Ryan

Semester 2 2022

Contents

In this practical, we are going to look at predicting the relationship between City Miles per Gallon and Engine Displacement, while taking into account the Drive (type).

Load data

The mpg data is included in the `tidyverse` package. To load the data:

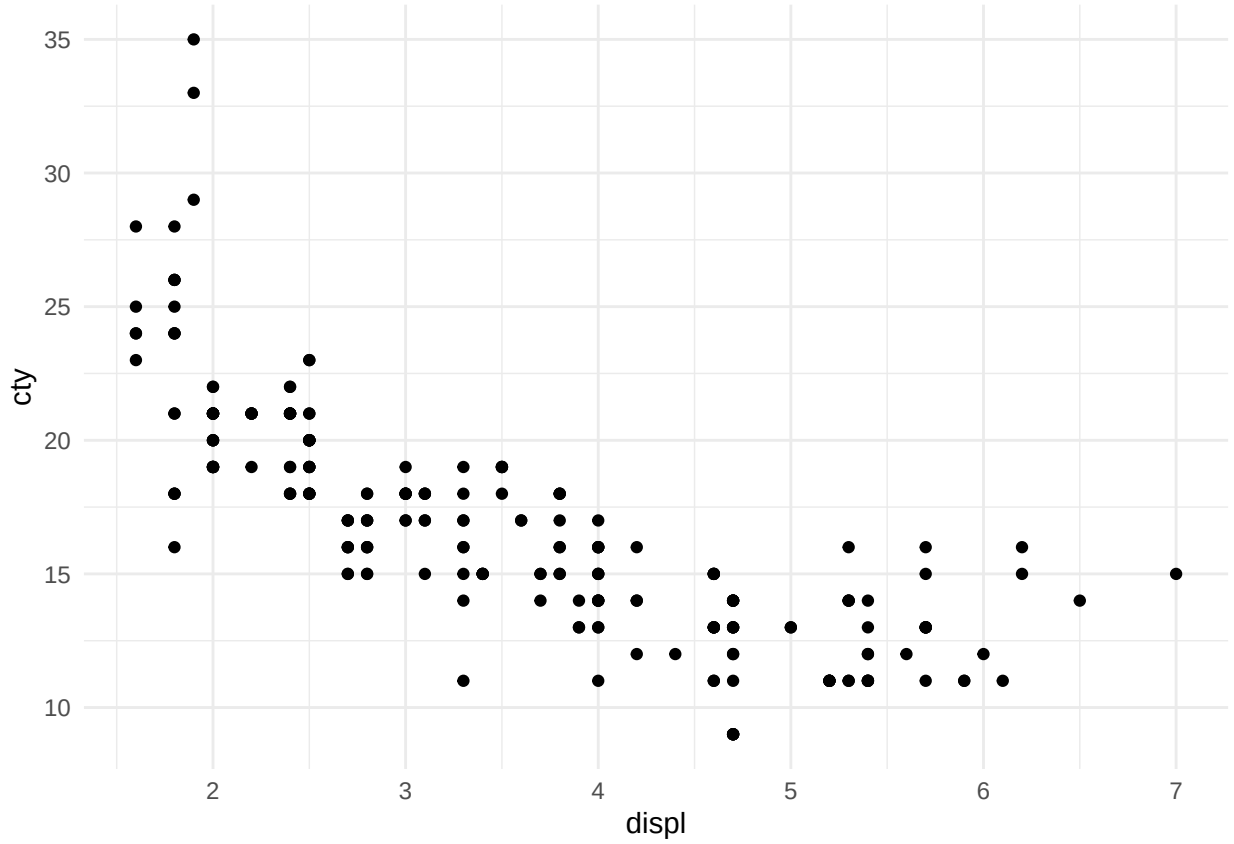
```
library(tidyverse)
data(mpg)
```

Task 1: Primary examination of the data

Quiz Questions

1. For the following scatter-plot, describe the relationship between Displacement and City Miles per Gallon.

```
mpg %>%
  ggplot(aes(displ, cty)) +
  geom_point()
```



2. Produce a side-by-side box plot of `cty` for each level of `drv`. Does there appear to be a difference in City Miles per Gallon for each different Drive type?

3. Produce a scatter-plot of Displacement vs City miles per gallon with different colours and lines for each Drive (type). Describe the relationship between Displacement and City Miles per Gallon accounting for Drive (type).

4. Do you think that there is an interaction between Drive (type) and Displacement in predicting City Miles per Gallon?

Task 2: ANOVA

Quiz Questions

5. By fitting the linear model `cty ~ drv`, perform an ANOVA to test the hypothesis

$$H_0 : \mu_f = \mu_r = \mu_4$$

where μ_f is the true mean cty for the front wheel drive cars, μ_r is the true mean cty for the rear wheel drive cars, and μ_4 is the true mean cty for the 4-wheel drive cars. You must report:

- The observed value of the test statistic
- The degrees of freedom for the reference F-distribution
- The associated p-value

Task 3: Identical regression lines

Quiz Questions

6. For the following observations, write down the design matrix for the identical regression of City Miles per Gallon on Displacement.

drv	displ	cty
4	1.8	18
f	1.8	18
r	5.3	14

Quiz Questions

7. Fit a linear model for cty vs displ with identical regression lines for drv, that is, fit the model

$$\text{cty}_i = \beta_0 + \beta_1 \text{displ}_i + \varepsilon_i \,.$$

   Using the model output, write down the line of best fit.

Task 4: Parallel regression lines

Quiz Questions

8. For the following observations, write down the design matrix for the parallel regression of City Miles per Gallon on Displacement and Drive (type).

drv	displ	cty
4	1.8	18
f	1.8	18
r	5.3	14

9. Fit a linear model for the parallel regression of City Miles per Gallon on Displacement and Drive (type), that is, fit the model

$$\text{cty}_i = \beta_0 + \beta_1 \text{displ}_i + \beta_2 \text{drv(f)}_i + \beta_3 \text{drv(r)}_i + \varepsilon_i \,,$$

where

$$\text{drv(f)}_i = \begin{cases} 1 & \text{if subject i is a front-wheel drive,} \\ 0 & \text{otherwise,} \end{cases} \qquad \text{drv(r)}_i = \begin{cases} 1 & \text{if subject i is a rear-wheel drive,} \\ 0 & \text{otherwise.} \end{cases}$$

Using the model output, write down the line of best fit for each Drive (type).

Task 5: Separate regression lines

Quiz Questions

10. For the following observations, write down the design matrix for the separate regression of City Miles per Gallon on Displacement and Drive.

drv	displ	cty
4	1.8	18
f	1.8	18
r	5.3	14

11. Fit a linear model for the separate regression of City Miles per Gallon on Displacement and Drive, that is, fit the model

$$\text{cty}_i = \beta_0 + \beta_1 \text{displ}_i + \beta_2 \text{drv(f)}_i + \beta_3 \text{drv(r)}_i + \beta_4 \text{displ}_i \times \text{drv(f)}_i + \beta_5 \text{displ}_i \times \text{drv(r)}_i + \varepsilon_i \,,$$

where drv(f)_i and drv(r)_i are defined as above. Using the model output, write down the line of best fit for each Drive type.

Task 6: Model Selection

Quiz Questions

12. Use AIC or BIC to decide on the most appropriate model. Use your final model to predict the mean City Miles per Gallon for a four-wheel drive with a displacement of 4 litres.

Task 7: Model contrasts (OPTIONAL)

13. Calculate the 95% confidence interval for the **difference** in mean City Miles per Gallon for the rear-wheel and front-wheel drive (i.e. drv(r) - drv(f)) by:
    a. Loading the `emmeans` package.
    b. Creating an `emmeans` object that looks at Drive type for the separate regressions model.
    c. Creating a *contrast* to look at the hypothesis drv(r) - drv(f) (Hint, what is your contrast vector going to be?)
    d. Use the `confint` function to obtain your confidence intervals.