MATH501 Modelling and Analytics for Data Science Coursework

Dr Julian Stander

Module Taught in Semester 2, Academic Year 2024/25

1 Coursework Information

Please read the following points before attempting the coursework:

The deadline for this assignment is **2pm** on **Friday, 2nd May, 2025**. You should submit your work through the MATH501 Modelling and Analytics for Data Science DLE site.

- **This is a group coursework. Please work in self-assigned groups of up to four people.** You will need to register your group on the DLE.
- Each member of the group will receive the same mark, unless any member chooses to make use of the Peer Assessment option. You should keep notes of all your group meetings to use as evidence in case you choose to make use of the Peer Assessment option. If you wish to make use of this option, you will need to make an appointment to see the Module Leader **Dr Julian Stander** by sending an email to J.Stander@plymouth.ac.uk.
- If you are experiencing problems with your group, please talk to the Module Leader **Dr Julian Stander** as soon as you can, and certainly not after the coursework has been submitted, when it will be **too late** to do anything.
- Each member of the group can expect to spend up to around 70 hours on this coursework and associated study.
- You can work on your own, but are strongly advised not to do so.
- In the event of your having valid extenuating circumstances affecting this submission, the maximum permitted extension to the deadline will be **5 WORKING DAYS**.
- Your coursework must be submitted electronically as a pdf file[1] using the online submission facility on the MATH501 Modelling and Analytics for Data Science DLE by the deadline.
- Your submission should be anonymous, so that it can be marked anonymously.

[1]This is our standard link to help about scanning, if you need it.

- Please ensure you are familiar with the regulations regarding Academic Offences. Please also see here for the definition of plagiarism and here for the academic regulations and procedures.
- Academic offences occur when activity is undertaken which could confer an unfair advantage to any candidate(s) in assessment. The University recognises the following (including any attempt to carry out the actions described) as academic offences, regardless of intent:

  a. *Plagiarism, which is copying or paraphrasing of other people's work or ideas into a submitted assessment without full acknowledgement. More information on plagiarism is available here.*

  b. *Collusion, which is unauthorised collaboration of students (or others) in producing a submitted assessment. The offence of collusion occurs if a student copies any part of another student's work, or allows their own work to be copied. Collusion also occurs if other people contribute significantly to work that a student submits as their own.*

- Please read the information on the MATH501 Modelling and Analytics for Data Science DLE about the **Use of AI**.
- Please remember to **GIVE CREDIT WHERE CREDIT IS DUE**, by providing proper references.
- Computer problems will not be considered as valid Extenuating Circumstances. Consequently, you should give due consideration to your personal time management to ensure that your work is submitted on time.
- For general information about Extenuating Circumstances, please see here.
- For the 'Extenuating Circumstances Policy and Procedures', please see here.
- This assignment counts for **all** of your final mark on this module (please note that at the moment the Machine Learning part of this coursework is not avaiable). Marks will be assigned according to the marking grid on page 4.

We now state the relevant MATH501 Modelling and Analytics for Data Science Assessed Learning Outcomes (ALOs) for this assignment.

At the end of the module the learner will be expected to be able to:

ALO1 display an in-depth understanding of a broad range of up-to-date modelling and analytics techniques for Data Science and a critical awareness of their limitations;

ALO2 critically choose and evaluate appropriate modelling or analytics techniques in new and complex practical situations to yield insight and innovation;

ALO3 present results professionally and systematically to technical and non-technical audiences.

You should keep these ALOs in mind when working on this coursework.

2 What You Need to Do

You will need to produce a report of your work following the instructions below, taking account of the marking grid on page 4.

Your report should contain well presented and annotated **R** or `Stan` code for **all** of your analyses.

The very generous **total page limit for everything** (First Task, Second Task, etc) **is thirty five pages, including R code and figures.** You should use clearly readable fonts and maintain sensible margins and line spacing throughout. Your work should not look squashed. Please see page 16 for more details. **Please do not submit an additional appendix as it will not be considered.** Reports that contain irrelevant or uninteresting discussion or **R** code will be penalized. Please note that **thirty five pages** is the limit and should not be considered as the target. Well written, concise reports will receive high credit.

As a general rule, **it is not necessary to repeat figures or R code that are very similar, for example.**

Stars indicate the approximate relative importance of the individual parts, with more stars indicating that the part is more important. They are included as an indicative guide only.

Please include part numbers as subheadings.

MATH501 Modelling and Analytics for Data Science: Coursework (CW) Marking Grid

Assessment Area	Maximum Mark	Awarded Mark	Feedback
CW First Task: Statistical Modelling: Clear and concise explanation and presentation of modelling results. Depth of understanding and quality of reporting of insights gained. Correct, well written and clearly annotated **R** and `Stan` code.	45		
CW Second Task: Machine Learning: Clear and concise explanation and presentation of Machine Learning results. Depth of understanding and quality of reporting of insights gained. Correct, well written and clearly annotated **R** code.	35		
CW Report: Well structured and presented discussion. Inferential and computational results displayed clearly, concisely, systematically and **professionally**. Proper credit given to others. Correct spelling and grammar. This is important for employability.	20		
CW Total[2]	100		

[2]This mark is provisional. Like all marks it is considered by an External Examiner and an Assessment Panel.

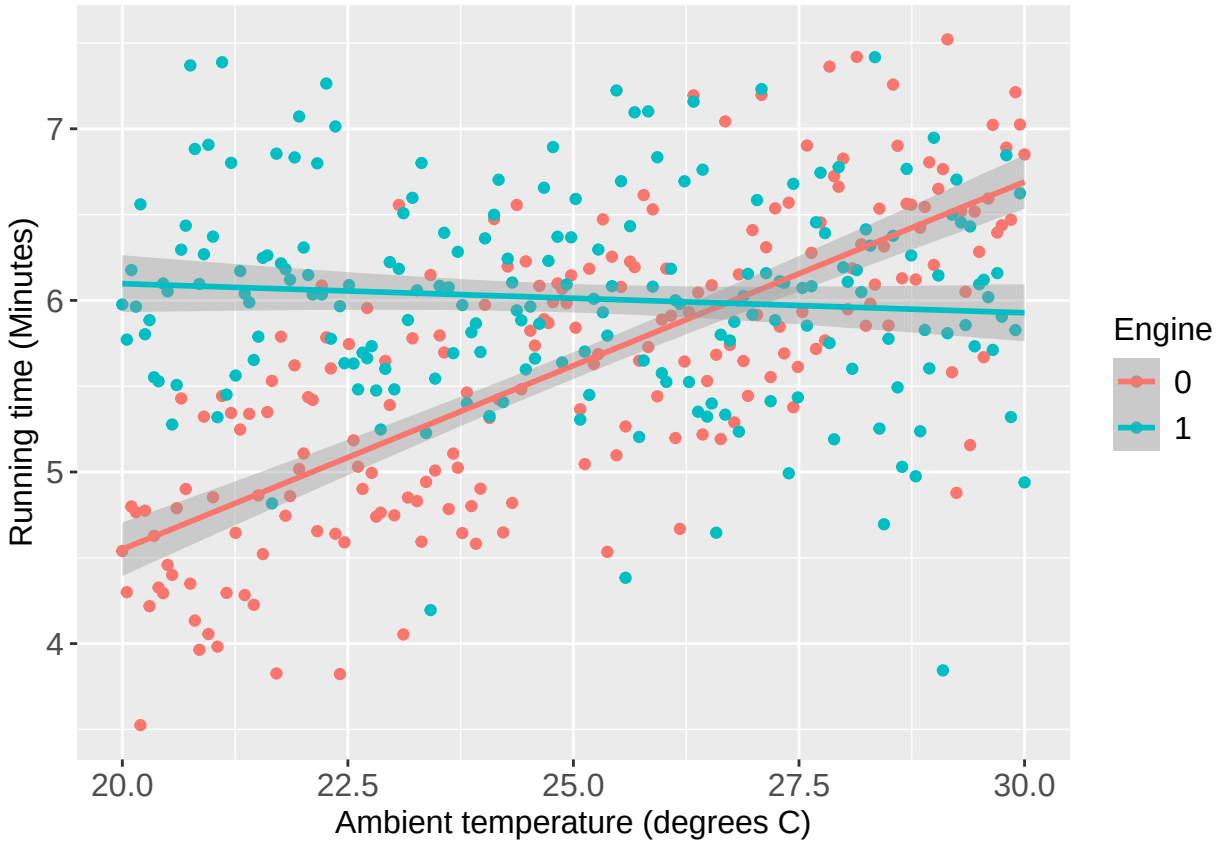
3 CW First Task: The General Linear Model – The Analysis of Covariance

3.1 Introduction

The fuel consumption of a car engine is tested by measuring the time that the engine runs at constant speed on a litre of standard fuel. Two engines are compared for fuel consumption: Engine 0 is a standard type, while Engine 1 is a new type under experimentation. For each engine a series of tests is performed at various fixed ambient temperatures, meaning that we are dealing with data from a designed experiment. The data are available from the file `MATH501_2024_25_fuel_consumption_data.csv` that is supplied.

Task 1.1*: Write and present code to read these data into **R**. Using the **R** package `dplyr` (part of the `tidyverse`), or otherwise, transform the variable Engine into a **factor** with levels 0 and 1.

Task 1.2*: Using the **R** package `ggplot2`, or otherwise, produce or improve on the following plot which illustrates the dependencies of running time on ambient temperature for the two engines. Provide a very brief description of your plot.



Task 1.3*:**

Consider the following 'analysis of covariance' model

$$\left.\begin{array}{l} y_i = (\beta_0^0 + \delta_0 I[\text{Engine}_i = 1]) + (\beta_1^0 + \delta_1 I[\text{Engine}_i = 1])\, x_i + \epsilon_i,\ i = 1, \ldots, n \\ \epsilon_i \sim N(0, \sigma^2) \text{ independently,} \end{array}\right\} \quad (1)$$

in which, for the i-th observation, y_i is the running time (minutes), x_i is the ambient temperature (degrees C), n is the overall sample size, $I[\text{Engine}_i = 1] = 1$ if $\text{Engine}_i = 1$, and 0 otherwise, and σ is standard deviation.

Now, perform the following sub-tasks:

- Explain carefully what δ_0 and δ_1 represent in model (1).
- Fit model (1) in the frequentist framework. Report and discuss your results briefly, but carefully. Using general notation, such a model can be fitted by means of code similar to

  ```
  lm(y ~ x * Engine)
  ```

  in which * means 'interaction'.
- State 95% confidence intervals for δ_0 and δ_1.
- Find a point (i.e. single number) estimate for $\beta_0^1 = \beta_0^0 + \delta_0$ and for $\beta_1^1 = \beta_1^0 + \delta_1$, and report your results carefully.
- Provide a point estimate for σ.

To help you with some of the above, the following **R** code illustrates how to extract parts of an **R** vector or matrix. You are not obliged to use code like this in your answer, but it may help.

```
v <- c(1, 3, 2, 5)
v
```

```
## [1] 1 3 2 5
```

```
v[c(2, 4)]
```

```
## [1] 3 5
```

```
#
M <- matrix(c(1, 3, 2, 5, 7, 8, 6, 1),
            ncol = 2,
            byrow = TRUE)
M
```

```
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    5
## [3,]    7    8
## [4,]    6    1
```

```
M[c(2, 4),]
```

```
##      [,1] [,2]
## [1,]    2    5
## [2,]    6    1
```

Task 1.4****:** Discuss briefly whether the studentized residuals associated with model (1) are consistent with the normal error assumption $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, independently.

Task 1.5****:** Provide a point estimate and a 95% confidence interval for the mean running time and a 95% prediction interval for running time for both engines when the ambient temperature is 18 degrees C and when it is 28 degrees C.

Explain what a 95% confidence interval for mean running time and a 95% prediction interval for running time represent.

Provide a brief critical discussion about what your confidence intervals tell you about engine performance.

Please note that to specify the values of a factor, you may need to use inverted commas, e.g. `"0"`.

Now that you have performed a frequentist analysis of the data provided in the file `MATH501_2024_25_fuel_consumption_data.csv`, you are going to perform a somewhat more sophisticated **Bayesian analysis**. This will involve producing inference about a quantity that is hard to learn about in the frequentist framework, so providing another example of the advantage of working in the Bayesian framework.

Please note that for the Bayesian analysis required for this coursework it is completely acceptable to perform and submit the main Bayesian computational tasks using **one** collection of `Stan` code that you build up as the tasks proceed.

We will be concentrating on the following model, which allows different error variances:

$$
\left.
\begin{array}{l}
y_j = \beta_0^0 + \beta_1^0 x_j + \epsilon_j^0, \text{ for } j \text{ such that } \text{Engine}_j = 0 \\
\epsilon_j^0 \sim N(0, \text{variance } \sigma_0^2) \text{ independently} \\
\\
y_j = \beta_0^1 + \beta_1^1 x_j + \epsilon_j^1, \text{ for } j \text{ such that } \text{Engine}_j = 1 \\
\epsilon_j^1 \sim N(0, \text{precision } \sigma_1^2) \text{ independently} \\
\\
\text{Priors:} \\
\\
\beta_0^0 \sim N(0, \text{variance } 100^2) \\
\beta_1^0 \sim N(0, \text{variance } 100^2) \\
\beta_0^1 \sim N(0, \text{variance } 100^2) \\
\beta_1^1 \sim N(0, \text{variance } 100^2) \\
\\
\sigma^0 \sim \text{Cauchy}(1.0, 10.0) \text{ positive part} \\
\sigma^1 \sim \text{Cauchy}(1.0, 10.0) \text{ positive part}
\end{array}
\right\} \quad (2)
$$

Task 1.6****: Write `Stan` code to implement model (2). Run your code. Specify 2 chains, 5000 (or more) McMC iterations, with a burn-in of length 2500 (or more) iterations and a thinning interval of 1, meaning that there is no thinning. Present your code in your report.

Please note: for the development phase of my code, I used 2000 McMC iterations and a burn-in of length 1000 to avoid having to wait too long.

Now, perform the following sub-tasks:

- Produce traceplots of β_0^0, β_1^0, β_0^1, β_1^1, σ^0 and σ^1. Comment very briefly on Markov chain Monte Carlo convergence.
- Produce well designed density estimates of β_0^0 and β_0^1. Produce well designed density estimates of β_1^0 and β_1^1. Produce well designed density estimates of σ_0 and σ_1. Explain briefly what is being shown in these plots. Is it sensible to assume that $\sigma_0 = \sigma_1$? Versions of these plots are included here, just so you can check that your code is working properly.

Here is some help (that you are not obliged to follow):

- When I code, I start simply and then build up code complexity. For example, one could start with code that performs Bayesian simple linear regression; get the code working on the data being considered; then modify the code.
- I specified my data model using a loop as

```
for(i in 1:n){
//
Running_time[i] ~ normal((beta_0_0 * (1 - Engine[i]) + beta_0_1 * Engine[i]) +
    (beta_1_0 * (1 - Engine[i]) + beta_1_1 * Engine[i]) * Ambient_temperature[i],
      sigma_0 * (1 - Engine[i]) + sigma_1 * Engine[i]);
}
```

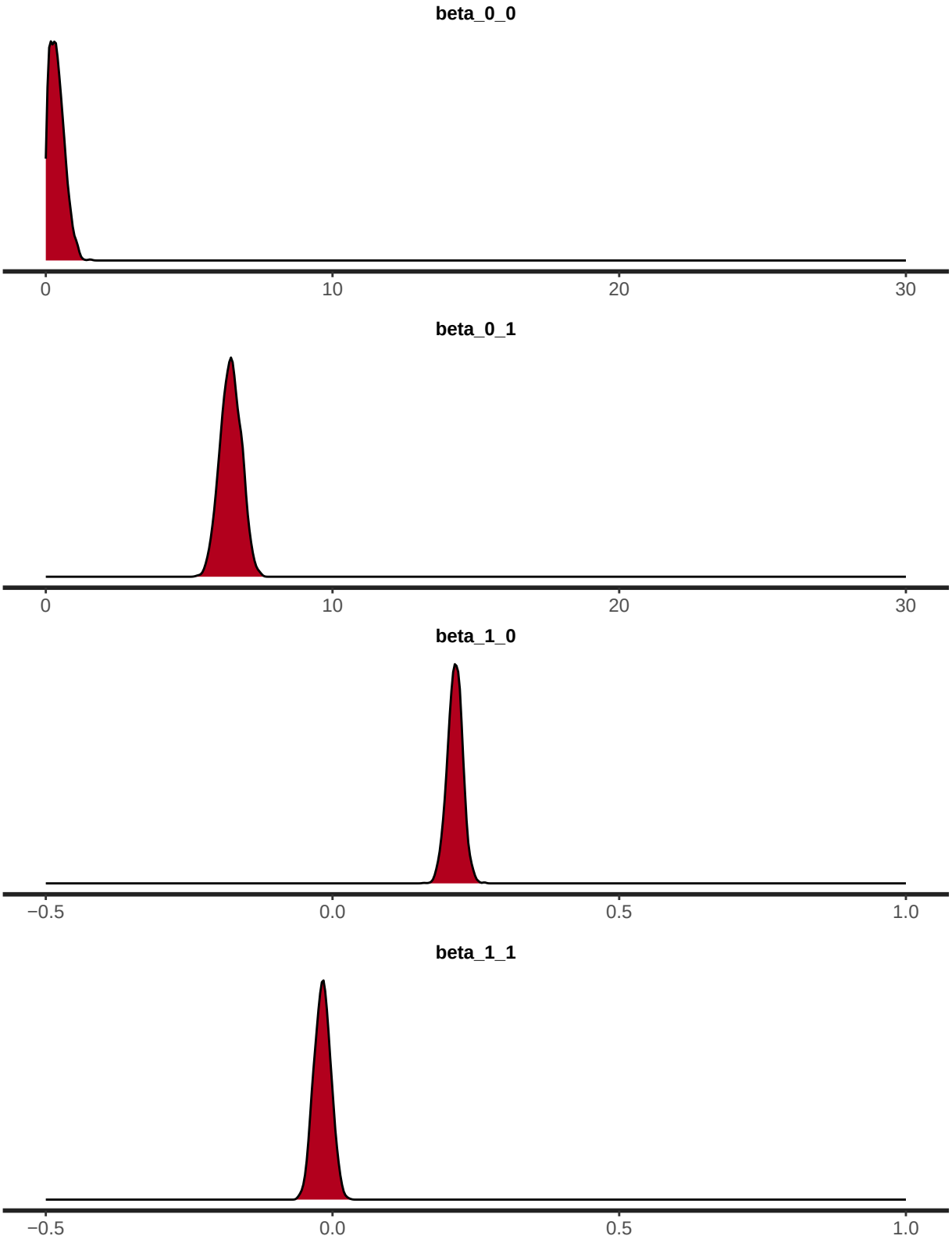
  although this is not the only way of specifying the model.

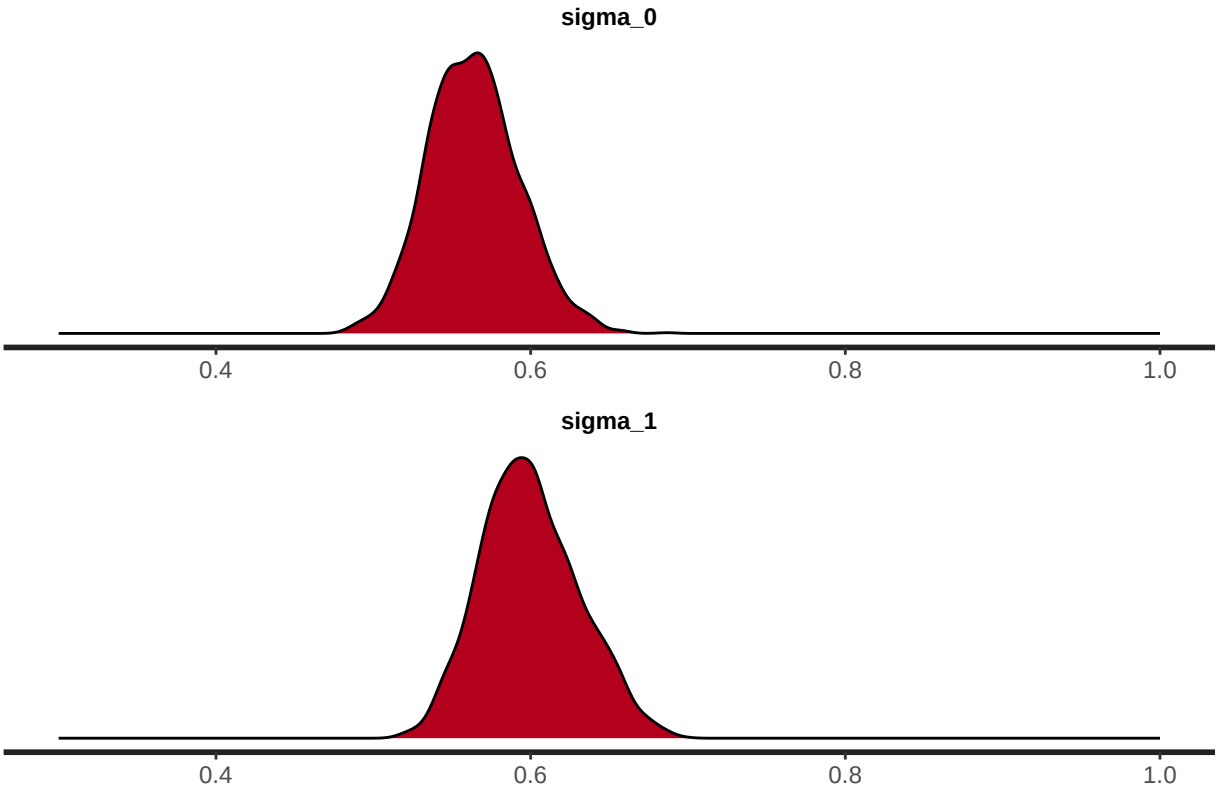
- When we want to look at the results of a `Stan` model fit, we can use the `print` function, which has an argument `pars`. Here is a simplified example:

```
print(fit, # Stan model fit
      pars = c("beta_0_0",
               "beta_1_0"))
```

  Please note that `print` also has a `probs` argument.

- If you need to, you can control the x-axis of a `stan_dens` plot using `+ xlim(1.5, 4)`, for example. You may need to load `ggplot2`.


beta_0_0
0
10
20
30
beta_0_1
0
10
20
30
beta_1_0
−0.5
0.0
0.5
1.0
beta_1_1
−0.5
0.0
0.5
1.0


sigma_0
0.4
0.6
0.8
1.0
sigma_1
0.4
0.6
0.8
1.0

Task 1.7*****:** It is of interest to perform inference about the difference in regression line values $(\beta_0^1 + \beta_1^1 x) - (\beta_0^0 + \beta_1^0 x)$ when the ambient temperature x is 23 and when it is 29 degrees C.

Provide the numerical values of 90% credible intervals for these two quantities. Explain precisely the meaning of one of these two intervals.

Task 1.8*****:** Find a mathematical expression for the x value, $x^{\text{intersect}}$, at which the two lines

$$y = \beta_0^1 + \beta_1^1 x \text{ and } y = \beta_0^0 + \beta_1^0 x$$

intersect. Include this expression in your report.

Now, perform the following sub-tasks:

- Provide the numerical values of a 95% credible interval for $x^{\text{intersect}}$.
- Produce a density estimate for $x^{\text{intersect}}$ and add vertical lines at 27 and 27.25 degrees C. A version of this plot is supplied on page 14.
- Find an approximate value of the posterior probability

$$\Pr\left(27 < x^{\text{intersect}} < 27.25 \,|\, \text{Data}\right).$$

Please note that it would be very difficult to make inference about $x^{\text{intersect}}$ in the frequentist framework!

Here is some help:

- It is possible to add vertical lines to the plot produced by `stan_dens` using `+ geom_vline(xintercept = ...)` where appropriate values are specified in the usual **R** way. You may need to load `ggplot2`.
- Sampled parameter values can be extracted from `Stan` output using the `extract` function. Here is a simplified example:

```
extract(fit, # Stan model fit
        pars = "Ambient_temperature_intersection",
        inc_warmup = FALSE)
```

- Please consider the following code:

```
# Generate and display some data for illustration
#
z <- rnorm(10000)
#
hist(z,
     freq = FALSE)
#
abline(v = c(-1.96, 1.96),
       lwd = 2)
```


```
#
# What's happening here?
#
mean(-1.96 < z & z < 1.96)
```

```
## [1] 0.948
```

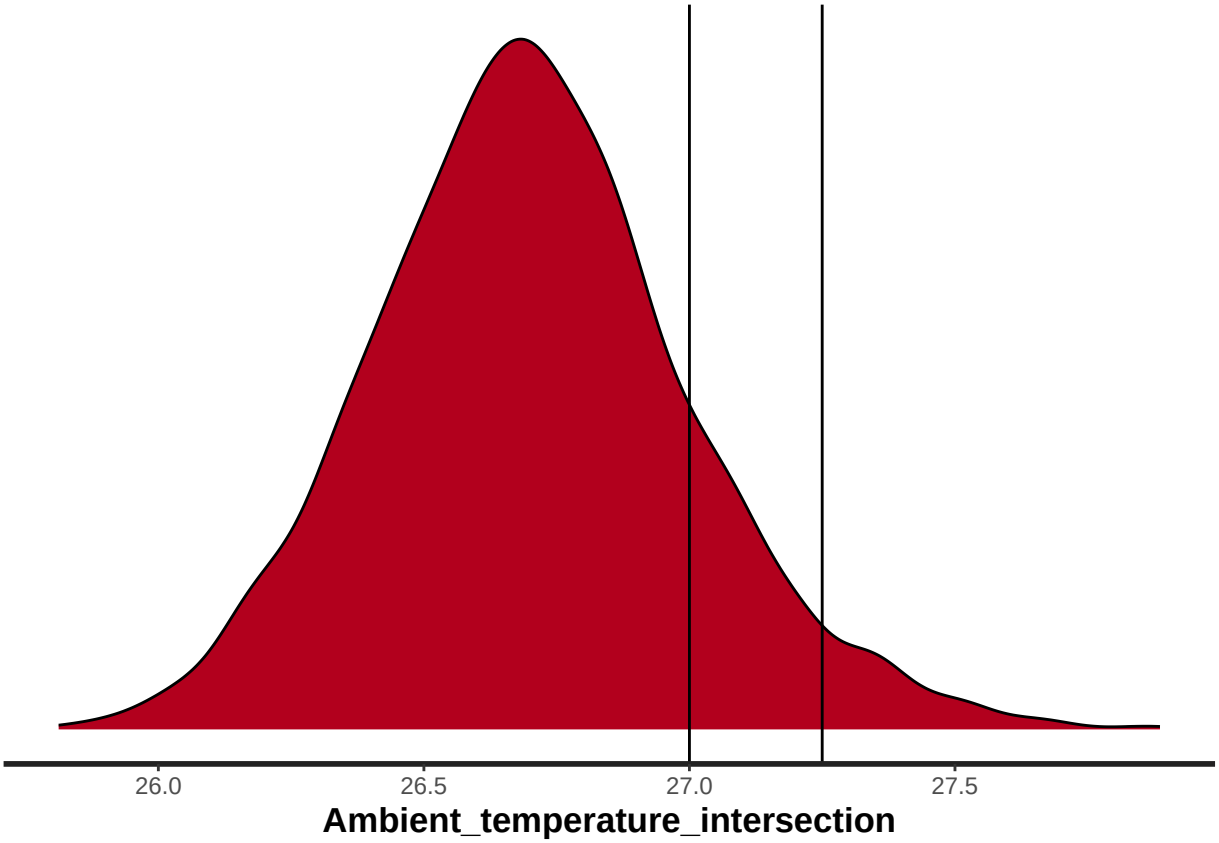
```
#
head(-1.96 < z & z < 1.96)
```

```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE
```

```
#
logical_test <- c(TRUE, FALSE, TRUE)
#
mean(logical_test)
```

```
## [1] 0.6666667
```

Here is the density estimate mentioned above:



Task 1.9*: Write a paragraph of no more than 200 words in which you explain to a graduate engineer who has not studied MATH501 Modelling and Analytics for Data Science how the fuel consumption for Engine 0 and Engine 1 depends on ambient temperature. Please state the number of words that you use. A concise well written paragraph will receive high credit. You may use bullet points.

4 CW Second Task: Machine Learning

This will follow...

5 Report Production

The ability to report technical results clearly and concisely is an important skill that is highly relevant to employability.

You should write a report that, as a minimum, discusses in detail your work and analyses, and contains well-presented and clearly annotated **R** and `Stan` code, as appropriate.

You are advised, but not obliged to use R Markdown or Quarto. We will certainly help you to get started with R Markdown or Quarto, if you have not yet met them. You can use LaTeX from within R Markdown or Quarto, but you are under no obligation to use LaTeX.

You should use clearly readable fonts and maintain sensible margins and line spacing throughout. Your work should not look squashed. The default settings of R Markdown or Quarto are completely acceptable.

6 What You Need to Submit

You need to submit the following file electronically using the DLE:

- A Portable Document Format file `MATH501_Coursework_January_Cohort_2024_2025.pdf` containing your report.

If anything is unclear, you should contact J.Stander@plymouth.ac.uk **without delay**.