# Bayes Assignment 2 of 2025

Sean van der Merwe

## Instructions

Your group (maximum 3 people) task with this assignment is to constructively criticise the hypothetical analysis below. For your convenience the statements in the analysis are numbered in square brackets. If it is not clear which statement is being criticised no marks will be awarded.

Marks will be given as follows:

1. 10 marks for following the instructions. This includes submitting your own version of the group document, with your own name and student number at the top (plus the obvious things like the date and the assignment name). List your group mates' names underneath, with a short statement as to each of their contributions to the project (according to you).
2. For each valid criticism that is not substantiated / not logically supported you get +2.
3. For each valid criticism that is supported by logical argument you get +4.
4. For each valid criticism that is supported by logical argument with relevant citation(s) you get +8. Note that citations do not have to all be formal, the *occasional* informal citation will be accepted.
5. Except: small issues, such as code mistakes, that do not warrant citations will only get +4. So statistical issues count more than technical or detail issues.
6. For each invalid criticism or silly nitpick you get -4! (Don't waste time on packages or coding styles.)
7. Up to +8 marks can be earned for reporting the correct regression formula, that is, the linear model specification used to generate the data.
8. For sloppy work you get as much as -40. Examples include not clearly identifying the target of a criticism, excessive spelling/grammar mistakes, wasting space, not spacing at all, etc.
9. For each criticism where the wording matches that of another group, or the wording of a paper, you get -40.
10. If AI tools are used then indicate that clearly with quotations and summaries of the prompting approach. Simple AI copy-paste will get you -40 per suspected instance.
11. The maximum mark is 100 and the minimum mark is 1 (IF your name and student number are correct and you submit in time). *Tip: since it must be possible to get 100, there must be lots of serious statistical errors. Some statements may have multiple issues.*

In order to figure out all the errors you will need to run additional code and try things for yourself. For this reason the data is provided in *BayesAssignment2of2025.xlsx* and the code is provided in the erroneous report below.

## Analysis to criticise

## Read data

First we read in the data from Excel:

```
d <- openxlsx::read.xlsx('BayesAssignment2of2025.xlsx', 'Data')
names(d)

[1] "y"  "x1" "x2" "x3"
```

The client verbally informed us that *y* represents the average profit in their shoe shop (in R/pair) over a random 2 hour period, split by sales person. *x1* represents the average experience level (in months) of that sales person. *x2* is what they paid for diesel for their generator (R/l) and *x3* is whether there was load shedding at that time.

[1]The client mentioned that they took a random sample from the data set for us, so that we [2]“don’t need to worry about annoying factors such as Date or Time”, and so that [3]the data set can be small and easy to email. [4]We thanked them for their consideration.

## Clean data

[5]First we generate a simple summary to check for obvious issues.

```
summary(d)

       y                x1               x2              x3       
 Min.   : 351.0   Min.   : 3.900   Min.   :10.00   Min.   :0.00  
 1st Qu.: 588.0   1st Qu.: 6.500   1st Qu.:13.00   1st Qu.:0.00  
 Median : 617.5   Median : 7.750   Median :15.00   Median :0.00  
 Mean   : 604.0   Mean   : 7.790   Mean   :14.95   Mean   :0.15  
 3rd Qu.: 641.2   3rd Qu.: 9.225   3rd Qu.:17.00   3rd Qu.:0.00  
 Max.   :1152.0   Max.   :11.800   Max.   :20.00   Max.   :1.00  
```

We see a big *y* value. [6]We calculate a statistical boundary of $\bar{y} + 3\hat{\sigma}_y$ and see if any values are larger than that:

```
d$y[d$y > (mean(d$y) + 3*sd(d$y))]

[1] 1152
```

[7]We see that there is one large value, which we now remove as an outlier.

```
d <- d |> subset(d$y <= (mean(d$y) + 3*sd(d$y)))
```

## Check assumptions

[8]We do a normality test of *y* because we want to do a regression:

```
shapiro.test(d$y)


    Shapiro-Wilk normality test

data:  d$y
W = 0.8677, p-value = 1.224e-05
```

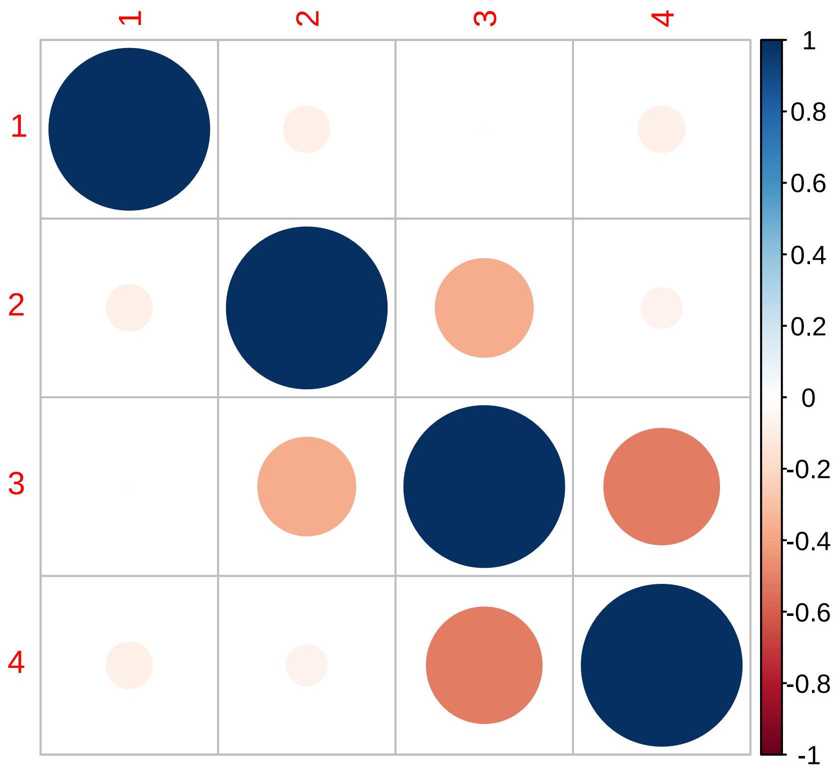
[9]Since the test rejects normality we cannot do a regression, [10]so we will try to use an ANCOVA instead. [11]For that to work we first check whether *y* differs by *x3*:

```
# [12]F test comparing group variances:
var.test(y ~ x3, data = d)


    F test to compare two variances

data:  y by x3
F = 0.57846, num df = 49, denom df = 8, p-value = 0.2307
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1518506 1.4251412
sample estimates:
ratio of variances
         0.5784644

# [13]F test rejects so we do unequal variance t-test, [14]otherwise we would do an
equal variance t-test, [15]or actually a z-test since the sample size is more than
30
t.test(y ~ x3, data = d)


    Welch Two Sample t-test

data:  y by x3
t = 7.6924, df = 9.7355, p-value = 1.935e-05
alternative hypothesis: true difference in means between group 0 and group 1 is not
equal to 0
95 percent confidence interval:
 135.6089 246.7777
sample estimates:
mean in group 0 mean in group 1
       623.8600        432.6667
```

[16]Since the *t*-test rejects we must include *x3* in the model as a main effect.

[17]We also want to check for auto-correlation. [18]The best way to do that is to cut our *x1* into segments and do a correlation matrix. [19]If any of the correlations are significant [20]then we know the residuals are correlated, [21]so then we can fit a mixed effects model.

```
n_obs <- nrow(d)
n_segments <- 4
segment_size <- ceiling(n_obs/n_segments)
segmented_matrix <- sapply(1:n_segments, \(i) {
  segment <- ((i-1)*segment_size + 1):min((i*segment_size), n_obs)
  c(d$x1[segment], rep(NA_real_, segment_size - length(segment)))
})
segmented_matrix |> cor(use = 'pairwise.complete.obs') |>
  corrplot::corrplot()
```



[22]None of the correlations are high so there are no problems.

## ANCOVA

[23]Luckily imbalanced samples is not a factor in ANCOVA, so [24]we don't need to worry about the way the errors are calculated.

```
aov(y ~ x1 + x2 + x3, data = d) |> summary()

            Df Sum Sq Mean Sq F value   Pr(>F)    
x1           1  46217   46217  15.594 0.000225 ***
x2           1     51      51   0.017 0.895807    
x3           1 252552  252552  85.216 9.04e-13 ***
Residuals   55 163002    2964                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[25]We see that *x2* is not significant. [26]We now redo the analysis with *x2* dropped.

```
aov(y ~ x1 + x3, data = d) |> summary()

            Df Sum Sq Mean Sq F value   Pr(>F)    
x1           1  46217   46217   15.74 0.000208 ***
x3           1 251232  251232   85.59 7.09e-13 ***
Residuals   56 164374    2935                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[27]Since we are using a stepwise approach we don't need to worry about over-fitting or parsimony.

[28] **This model is now ready for prediction.**

## Prediction

[29]We can now confidently make predictions of the shop's profit by plugging in new values into the formula above, regardless of what those values might be. [30]Since the model fit is significant we can be confident in the accuracy of our predictions.