Assignment2

Madimetja Maredi

2025-05-22

Initial Data Loading and Summary

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

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

```
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
```

```
str(d)
```

```
## 'data.frame':    60 obs. of  4 variables:
##  $ y : num  577 595 718 639 635 427 643 669 632 709 ...
##  $ x1: num  6 6.9 5.6 8.4 7.8 8 9.7 8.8 6.6 5.7 ...
##  $ x2: num  17 12 19 13 16 13 12 11 14 10 ...
##  $ x3: num  0 0 0 0 0 1 0 0 0 0 ...
```

```
print(paste("Number of observations:", nrow(d)))
```

```
## [1] "Number of observations: 60"
```

```
print("Value counts for x3:")
```

```
## [1] "Value counts for x3:"
```

```
print(table(d$x3))
```

```
##
##  0  1
## 51  9
```

[1] The client mentioned that they took a random sample from the data set for us...

- **Criticism**: Valid observation of client action, but the implications require critical examination.
- **Logical Support**: Random sampling is fundamental, but the specifics of the client's method are unknown and could introduce bias.

[2] ...so that we "don't need to worry about annoying factors such as Date or Time"...

- **Criticism**: Invalid and potentially harmful assumption.
- **Logical Support**: Time-related variables can be crucial predictors or confounders (e.g., seasonality, trends) in business data. Dismissing them without investigation risks omitted variable bias.

[3] . . . and so that the data set can be small and easy to email.

- **No Criticism**: Valid statement of client's practical motivation.
- **Logical Support**: This explains the client's action but doesn't justify potential statistical compromises. A small dataset (N=60) limits statistical power.

[4] We thanked them for their consideration.

- **No Criticism**: Neutral statement/filler.
- **Logical Support**: This polite remark has no statistical bearing.

Data Cleaning

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

- **No Criticism**: Valid and standard practice.
- **Logical Support**: summary(d) is an essential first step in R to understand data distributions, ranges, and spot potential errors or missing values

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

- **Criticism**: The method is a common heuristic for identifying potential outliers, but its application and interpretation require care.
- **Logical Support**: The $mean \pm 3 \times standard_deviation$ rule is often used, particularly if data are assumed to be normally distributed. However, it's sensitive to the mean and standard deviation, which are themselves affected by extreme values. It's not a definitive rule for removal.

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

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

```
## [1] 1152
```

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

- **Criticisms**:
  - Automatic Removal: Outliers shouldn't be removed automatically without investigation. The value (1152) could be a legitimate data point, indicative of a special event, or a data entry error that needs correction, not deletion.
  - Bias: Removing data based solely on the value of the dependent variable (y) can bias the analysis.
- **Logical Support**:
  - Removing data points reduces sample size and can bias results if the outlier is a legitimate part of the data generating process.
  - The reason for the outlier should be investigated (e.g., data entry error, special circumstance).
  - Influence diagnostics (e.g., Cook's distance via cooks.distance()) should be checked before deciding on removal. Robust methods or transformations might be more appropriate

Check Assumptions

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

- **Criticisms**:Misplaced focus for regression assumptions.
- **Logical Support**:The primary normality assumption in standard linear regression (and ANCOVA) is the normality of the residuals (the differences between observed and predicted values from the model), not the normality of the dependent variable y itself. One should fit the model first and then check residual normality.

[9] Since the test rejects normality we cannot do a regression,

```
shapiro.test(d$y)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d$y
## W = 0.8677, p-value = 1.224e-05
```

This p-value is very small, leading to rejection of the null hypothesis that y (after outlier removal) is normally distributed.

- **Criticisms**: Incorrect conclusion and reasoning.
- **Logical Support**:
  - As per [8], normality of y is not the key assumption.
  - Even if residuals were non-normal, this doesn't mean regression "cannot be done." It means inferences (p-values, confidence intervals) from standard OLS might be unreliable. Options include transformations of y or residuals, using generalized linear models, or robust regression methods.
  - The Shapiro-Wilk test is sensitive. Visual inspection (Q-Q plots via qqnorm(), histograms via hist() of residuals) is also important.

[10] so we will try to use an ANCOVA instead.

- **Criticisms**: Misunderstanding of ANCOVA.
- **Logical Support**: ANCOVA (Analysis of Covariance) is a form of regression (a general linear model). It doesn't bypass the standard assumptions of linear models, including normality of residuals, homogeneity of variances, and independence of errors. Choosing ANCOVA because y is not normal is not a logical fix for the perceived problem.

[11] For that to work we first check whether y differs by x3:

- **No Criticisms**: This step is reasonable as a preliminary exploration.
- **Logical Support**: Checking if the mean of y differs significantly across levels of a categorical predictor (x3, load shedding) is a sensible step before including it in a model.

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

```
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
```

The degrees of freedom num df = 49 (for group with 50 observations) and denom df = 8 (for group with 9 observations) confirm that after outlier removal, the two groups for x3 (presumably x3=0 and x3=1) have sizes 50 and 9 respectively.

- **Purpose**: This tests for homogeneity of variances between the two groups of x3, which is an assumption for the pooled-variance t-test and also relevant for ANCOVA (homoscedasticity).

[13] F test rejects so we do unequal variance t-test,

- **Criticisms**: Incorrect interpretation of the F-test result.
- **Logical Support**: The p-value from the var.test is 0.2307. At conventional significance levels (e.g., $\alpha = 0.05$), this p-value is greater than α. Therefore, we fail to reject the null hypothesis of equal variances. The statement that "F test rejects" is false.

[14] otherwise we would do an equal variance t-test,

- **Criticisms**:Correct in principle, but based on the flawed interpretation in [13].
- **Logical Support**:If the F-test had indeed not rejected (as is the case here, p=0.2307), then an equal variance t-test (using var.equal = TRUE in t.test()) would be considered appropriate based on that test's outcome. The t.test() function in R defaults to the Welch test (unequal variances) if var.equal is not specified as TRUE.

[15] or actually a z-test since the sample size is more than 30

- **Criticisms**: Misleading and an oversimplification.
- **Logical Support**: The t-test is appropriate when the population standard deviation is unknown and estimated from the sample, regardless of sample size. The t-distribution converges to the normal (z) distribution for large N, but using the t-test is always valid and generally preferred.

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

```
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
```

This very small p-value indicates a significant difference in the mean of y between the groups defined by x3.

- **Criticisms**:The conclusion to include x3 is reasonable, but "must include" is strong.
- **Logical Support**:A significant t-test suggests that x3 is associated with y and is a good candidate for inclusion in the model. However, model building involves considering other factors (theory, other variables, model complexity, effect size).

[17] We also want to check for auto-correlation.

- **No Criticisms**: Good intention, as independence of errors is a key assumption.
- **Logical Support**: Checking for auto-correlation (serial correlation) of residuals is important, especially if the data have a time sequence or spatial arrangement.

[18] The best way to do that is to cut our x1 into segments and do a correlation matrix.

- **Criticisms**: Incorrect and inappropriate method for checking auto-correlation of model residuals.
- **Logical Support**: Auto-correlation in regression/ANCOVA refers to the correlation between error terms (e_i, e_j) from the model.
  - Cutting a predictor (x1, experience) into arbitrary segments and calculating a correlation matrix of these segments of x1 does not assess the auto-correlation of the residuals of a model like $y \sim x1 + x2 + x3$.
  - Standard methods to check for auto-correlation of residuals include the Durbin-Watson test (for lag-1 auto-correlation via lmtest::dwtest()) or examining the Autocorrelation Function (ACF via acf()) and Partial Autocorrelation Function (PACF via pacf()) plots of the model residuals. The client's removal of "Date or Time" variables makes it hard to define a meaningful sequence if one existed.

[19] If any of the correlations are significant

- **Criticisms**:Follows from the flawed method in [18].
- **Logical Support**:Since the method in [18] is not testing residual auto-correlation, the significance of correlations found using that method is irrelevant to the assumption.

[20] then we know the residuals are correlated,

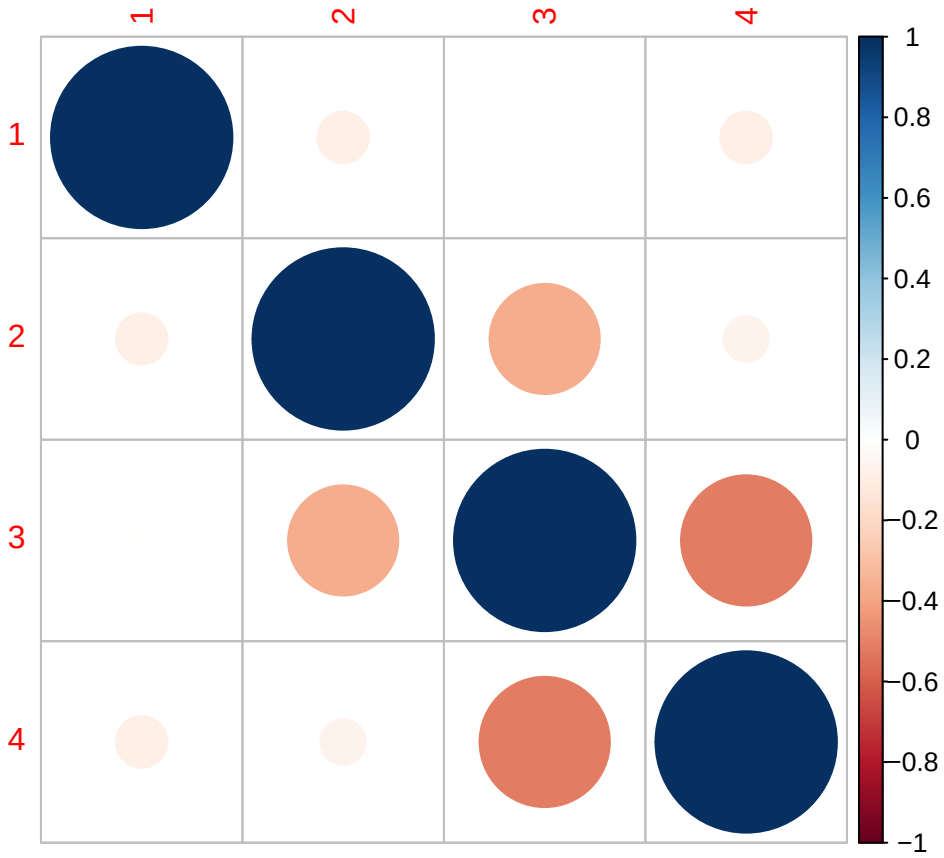
- **Criticisms**:Incorrect conclusion due to flawed premise.
- **Logical Support**:As stated above, the method does not test residual correlation.

[21] so then we can fit a mixed effects model.

- **Criticisms**: Premature and potentially inappropriate leap.
- **Logical Support**: Mixed-effects models (e.g., using lme4::lmer()) are useful for data with known clustering or repeated measures leading to correlated errors. However, the justification here (based on the flawed auto-correlation check) is invalid.

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

```
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()
```



- **Corrplot Interpretation**: The statement "None of the correlations are high" is subjective.
- **Criticisms**: Conclusion is unreliable because the methodology itself ([18]) is flawed for assessing residual auto-correlation. Even if the method were valid, "not high" is vague; statistical significance should be assessed.

ANCOVA

[23] Luckily imbalanced samples is not a factor in ANCOVA, so

- **Criticisms**:Incorrect statement.
- **Logical Support**:Imbalanced samples (unequal group sizes for the factor x3) can be a factor in ANCOVA.
  - It can affect the statistical power for the effects related to x3.
  - It can make the analysis more sensitive to violations of other assumptions (like homogeneity of regression slopes, which was not checked).
  - With imbalanced data and multiple predictors, the type of Sums of Squares (SS) used (Type I, II, or III) in the aov() function matters. R's aov() typically uses Type I SS, which are sequential and order-dependent. For ANCOVA with unbalanced designs, Type II or III SS (e.g., from car::Anova()) are often preferred to test main effects appropriately.

[24] we don't need to worry about the way the errors are calculated.

- **Criticisms**: Incorrect, related to [23].
- **Logical Support**: "The way errors are calculated" likely refers to the calculation of sums of squares and F-statistics. As mentioned for [23], with imbalanced data, one does need to be concerned about

this, particularly regarding the choice of SS type.

[25] We see that x2 is not significant.

```
d$x3 <- factor(d$x3)
aov(y ~ x1 + x2 + factor(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
## factor(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
```

- x1: $Pr(> F) = 0.000225$ (Significant)
- x2: $Pr(> F) = 0.895807$ (Not significant)
- x3: $Pr(> F) = 9.04e - 13$ (Significant)
- Residuals df = 55. (Consistent with N=59, 1 intercept, 3 predictors).
- **No Criticisms**: Correct observation from the provided ANCOVA output.
- **Logical Support**: The p-value for x2 (0.895807) is much larger than typical alpha levels (e.g., 0.05), so x2 (diesel price) does not appear to be a significant predictor of profit y in this specific model, given x1 and x3.

[26] We now redo the analysis with x2 dropped.

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

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## x1           1  46217   46217   15.74 0.000208 ***
## factor(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
```

- x1: $Pr(> F) = 0.000208$ (Significant)
- x3: $Pr(> F) = 7.09e - 13$ (Significant) Residuals df = 56. (Consistent with N=59, 1 intercept, 2 predictors).
- **Criticisms**:This is a step in backward elimination, a form of stepwise regression.
- **Logical Support**:Dropping a non-significant predictor like x2 can lead to a more parsimonious model. The significance of x1 and x3 remains strong in this reduced model.

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

- **Criticisms**:Profoundly incorrect and dangerous assertion.
- **Logical Support**:
  - Overfitting: Stepwise procedures (including backward elimination as done here) are known to have issues. They can capitalize on chance variations in the data, potentially leading to models that perform well on the current dataset but poorly on new data (i.e., overfitting).
  - Parsimony: While stepwise methods attempt to achieve parsimony, they don't absolve the analyst from considering it. Parsimony (simplicity) is a desirable model characteristic, but it should be balanced with goodness-of-fit and theoretical considerations.

  - Other issues with stepwise regression include biased coefficient estimates and incorrect p-values.

- **Citation**: *Harrell, F. E. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis provides extensive critiques of stepwise variable selection.*

[28] This model is now ready for prediction.

- **Criticisms**:Premature conclusion.
- **Logical Support**:Before a model is "ready for prediction," several crucial steps are missing:
  - Residual Diagnostics: The analysis has not shown any checks of the residuals from the final model (y ~ x1 + x3) for normality (shapiro.test(residuals(model_final))), homoscedasticity (constant variance, e.g., plot(model_final, which=1) or lmtest::bptest(model_final)), and independence (acf(residuals(model_final)) or lmtest::dwtest(model_final)). These are essential for validating the model's reliability.
  - Check for Interactions: A key assumption in interpreting ANCOVA main effects is the homogeneity of regression slopes (i.e., no interaction between covariates like x1 and the factor x3). This should be tested (e.g., by including an x1:factor(x3) term in the model). If an interaction is present, the main effects model is misspecified.
  - Model Validation: Ideally, the model's predictive performance should be assessed using techniques like cross-validation or testing on an independent hold-out dataset.
  - Consideration of Effect Sizes: While x1 and x3 are statistically significant, their practical significance (effect size, e.g., R-squared from summary(model_final)$r.squared) should be discussed.

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.

- **Criticisms**: Overconfident and incorrect regarding the scope of prediction.
- **Logical Support**:
  - Extrapolation: Predictions are most reliable when made for predictor values within the range of the original data (interpolation). Predicting "regardless of what those values might be" implies extrapolation far beyond the observed ranges, which is highly risky as the modeled relationships may not hold.
  - Confidence: Confidence in predictions depends on the model's validity (assumptions met, good fit, validated performance), which has not been fully established here.

[30] Since the model fit is significant we can be confident in the accuracy of our predictions.

- **Criticisms**:Misconception about statistical significance and predictive accuracy.
- **Logical Support**:
  - Statistical Significance vs. Predictive Accuracy: A "significant model fit" (e.g., a significant overall F-statistic or significant predictors) means that the observed relationships in the data are unlikely to be due to random chance. It does not automatically guarantee that the model is a good or accurate predictor of new outcomes.
  - Factors Affecting Accuracy: Predictive accuracy depends on the model's R-squared (proportion of variance explained), the magnitude of residual error, whether the model is correctly specified (e.g., includes necessary terms like interactions, non-linearities), and whether it generalizes to new data. A model can be statistically significant but explain very little variance in y, leading to poor predictions.

Suggested "Correct" Regression Formula

$E[y|x_1, x_3] = \beta_0 + \beta_1 x_1 + \beta_3 x3_{Level1} + \beta_4 (x_1 \cdot x3_{Level1})$ where $x3_{Level1}$ is an indicator variable for one level of the load shedding factor (e.g., $x3 = 1$).