Assignment 2

Assignment discussion: 27 August 2025.
Deadline: 6 October 2025.

Part 1: Multiple Linear Analysis

Question 1

The dataset `concrete.rda` contains laboratory data with the following variables:

- `cement` - cement content [kg/m^3]
- `wcr` = water-to-cement ratio [-]
- `age` = curing age [days]
- `strength` = compressive strength [MPa]

We are typically interested in predicting the strength of concrete if we know the cement content, the w/c ratio and/or the curing age.

(a) Before fitting multiple linear regression models, it is good practice to prepare your data. Investigate histograms, marginal distributions of the variables and scatter plots of strength against each predictor. Comment on any trends and the likely need for variable transformation.

(b) Investigate multicollinearity among the predictors:

- Compute the Pearson correlation coefficients.
- Use an ellipse plot to visualise collinearity.
- Calculate Variance Inflation Factors (VIF).

Comment on your findings

(c) Fit the multiple regression model: `strength` $\sim$ `cement + wcr + age`. Comment on the output of this model, its adequacy of fit and appropriateness.

(d) Use variable selection methods (backward elimination, forward selection, AIC stepwise) to refine the model. Which model(s) would you recommend and why?

(e) Evaluate the predictive power of your final model using 5-fold cross-validation. Report the mean square prediction error (MSPE). How would the MSPE compare if you were to add one of the predictor variables that you have left out in (d), or if you were to omit one of the predictor variables that you have in your final model in (d), depending on your final answer?

(f) For a mix with `cement` = 350 kg/m^3, `wcr` = 0.5, `age` = 28 days, predict the compressive strength and compute a 95% prediction interval. Comment on whether this prediction is practically useful.

Question 2

The dataset energy.rda contains energy consumption data from 80 office buildings:

- energy - annual energy use intensity [kWh/m^2]
- area - floor area [m^2]
- occup - average occupancy [people]
- climate - climate zone (1 = warm, 2 = temperate, 3 = cold)
- glazing - facade glazing percentage [%]
- insulation - insulation thickness [mm]

(a) Investigate multicollinearity among predictors with correlation coefficients and VIF. Which variables are problematic and why?

(b) Fit an initial multiple regression model for energy using all predictors, and assess the linearity of each predictor. If non-linearity is evident, consider an appropriate transformation. Comment on the model output parameters, adequacy of fit, and appropriateness of fit.

(c) Apply variable selection (backward elimination, AIC) starting from your appropriately-transformed model in (b). Compare the results.

(d) Perform a 5-fold cross-validation and compute MSPE for both the full and the reduced model. Which performs better for prediction?

Question 3

Multiple linear regression theory questions.

Q 3.1

You fit a multiple regression model predicting compressive strength from cement content, water–cement ratio, and age. The regression summary shows that none of the individual predictors are significant, but the overall F-test indicates the model is significant. Which of the following is the most likely explanation?

A. The predictors are uncorrelated with the response.
B. Multicollinearity is present among the predictors.
C. The sample size is too large.
D. The regression model is non-linear.

Q 3.2

Which of the following statements about variable selection in multiple regression is correct?

A. A model with the lowest AIC will always have the highest R^2.
B. Stepwise selection guarantees the true model will be chosen.
C. Models with more variables always have better predictive performance.
D. Cross-validation can help compare models based on predictive accuracy.

Part 2: Analysis of Variance

Question 4

The dataset `timber.rda` contains bending stiffness results:

- `species` - timber species [`pine`, `gum`, `cedar`]
- `stiffness` - bending stiffness [$\mathrm{kN} \cdot \mathrm{m}^2$]

(a) Produce boxplots of stiffness by species. Comment on the variability and possible outliers.

(b) Perform a one-way ANOVA to test whether stiffness differs significantly between species.

(c) Perform pairwise two-sample t-tests (with multiple comparison correction) to identify which groups differ.

(d) Check assumptions of the ANOVA (residual diagnostics).

Question 5

The dataset `curing.rda` contains compressive strength results for concrete cylinders cured under two different methods:

- `method` - curing method [`water`, `air`]
- `strength` - compressive strength [MPa]

(a) Create side-by-side boxplots of strength for the two curing methods. Based on an initial visual inspection, do you expect the two curing methods to result in a significant difference in strength?

(b) Perform a two-sample t-test to compare mean strengths. State the null and alternative hypotheses clearly.

(c) Report the test statistic, p-value, and your conclusion.

(d) Comment on whether the difference is practically significant in addition to being statistically significant.

Question 6

ANOVA theory questions.

Q 6.1

You run a one-way ANOVA comparing the bending stiffness of three timber species. The ANOVA table gives p = 0.01. What is the correct conclusion?

A. All three species differ significantly in stiffness.
B. At least one species has a mean stiffness significantly different from the others.
C. There is no difference in mean stiffness among the species.
D. The probability that the null hypothesis is true is 1%.

Q 6.2

Which of the following is not an assumption of one-way ANOVA?

A. The populations have equal variances.
B. The samples are independent.
C. The populations are normally distributed.
D. The sample sizes must be equal.

Q 6.3

In one-way ANOVA, the F-statistic is defined as:

A. The ratio of between-group variance to within-group variance.
B. The ratio of sample means to population means.
C. The square root of the sum of squares between groups.
D. The ratio of total variance to between-group variance.

Q 6.4

Why is one-way ANOVA preferred over multiple two-sample t-tests when comparing more than two groups?

A. ANOVA is faster to compute.
B. ANOVA avoids increasing the risk of Type I error from multiple t-tests.
C. ANOVA gives the exact same result as running all possible t-tests.
D. ANOVA requires fewer assumptions than t-tests.