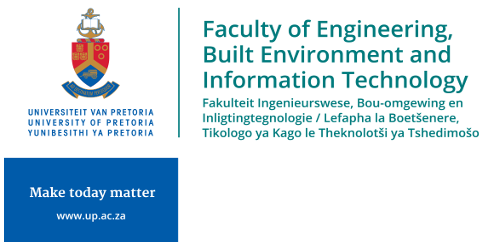
****

**DEPARTMENT OF CIVIL ENGINEERING**

**SHC 798**

**APPLIED STATISTICAL METHODS AND OPTIMISATION**

**Multiple Linear Regression & ANOVA**

**RICHARD LUBEGA**

*Full names*

**25585089**

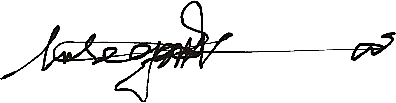
*Student number*

**2**

*Assignment*

# DECLARATION

1. I understand what plagiarism is and am aware of the University’s policy in this regard.
2. I declare that this submission is my original work. Wherever other people’s work has been used (either from a printed source, the internet or any other source) this has been properly acknowledged and referenced in accordance with departmental requirements.
3. I declare that I did not use ChatGPT or similar AI-based tools to prepare this report.
4. I have not used another student’s current or past written work to hand in as my own.
5. I have not allowed and will not allow anyone to copy my work to pass it off as his or her work.

Signature: **

Date: 25-08-2025

# TABLE OF CONTENTS

[DECLARATION 1](#_Toc210393348)

[TABLE OF CONTENTS 2](#_Toc210393349)

[LIST OF FIGURES 3](#_Toc210393350)

[LIST OF TABLES 3](#_Toc210393351)

[1 Part 1: Multiple Linear Analysis (MLR) 4](#_Toc210393352)

[1.1 Question 1 4](#_Toc210393353)

[1.1.1 Part a): Data Preparation 4](#_Toc210393354)

[1.1.2 Part b): Multicollinearity 4](#_Toc210393355)

[1.1.3 Part c) Model Output 4](#_Toc210393356)

[1.1.4 Part c): Variable Selection 6](#_Toc210393357)

[1.1.5 Part d): 5-fold Cross Validation & MSPE 6](#_Toc210393358)

[1.1.6 Part e): Prediction 6](#_Toc210393359)

[1.2 .Question 2 7](#_Toc210393360)

[1.2.1 Part a): Multicollinearity 7](#_Toc210393361)

[1.2.2 Part b): Model and Predictor Linearity 7](#_Toc210393362)

[1.2.3 Part c): Variable Selection 7](#_Toc210393363)

[1.2.4 Part d): 5-fold cross-validation & MSPE 7](#_Toc210393364)

[1.3 Question 3 8](#_Toc210393365)

[1.3.1 Q 3.1: MCQ Answer 8](#_Toc210393366)

[1.3.2 Q 3.2: MCQ Answer 8](#_Toc210393367)

[2 Part 2: Analysis of Variance (ANOVA) 9](#_Toc210393368)

[2.1 Question 4 9](#_Toc210393369)

[2.2 Question 5 9](#_Toc210393370)

[2.3 Question 6 9](#_Toc210393371)

[2.3.1 Q 6.1 9](#_Toc210393372)

[2.3.2 Q 6.2 9](#_Toc210393373)

[2.3.3 Q.6.3 9](#_Toc210393374)

[2.3.4 Q6.4 9](#_Toc210393375)

[3 Residual Write-Up 10](#_Toc210393376)

[REFERENCES 11](#_Toc210393377)

# LIST OF FIGURES

[**Figure 1**: CDF for Construction Cost 8](#_Toc206940718)

[**Figure 2**: CDF for Annual Trip Growth Rate 8](#_Toc206940719)

[**Figure 3**: CDF for Value of Travel Time Savings 8](#_Toc206940720)

[**Figure 4**: Computing Travel Savings Benefits 12](#_Toc206940721)

# LIST OF TABLES

[**Table 1**: The four project options (costs and benefits for 10 years) 4](#_Toc206940724)

[**Table 2**: Discounting and Computing the NPV, BCR and IRR 4](#_Toc206940725)

[**Table 3**: Obtaining the IRR values for the four projects 5](#_Toc206940726)

[**Table 4**: Ordering projects in increasing cost fashion 5](#_Toc206940727)

[**Table 5**: Incremental BCR analysis 6](#_Toc206940728)

[**Table 6**: Expected values for the input variables 7](#_Toc206940729)

[**Table 7**: The most likely BCR 7](#_Toc206940730)

[**Table 8**: Monte Carlo Simulation 10](#_Toc206940731)

[**Table 9**: Computing Traffic Growth for "Without" and "With" Bypass Project 12](#_Toc206940732)

[**Table 10**: Computing VOC Savings and VTTS for "With" and "Without" Bypass Route 14](#_Toc206940733)

[**Table 11**: Discounting to present worth terms 15](#_Toc206940734)

# Part 1: Multiple Linear Analysis (MLR)

## Question 1

Concrete Data

### Part a): Data Preparation

### Part b): Multicollinearity

#### Pearson correlation coefficients

#### Ellipse plot to visualise collinearity

correlogram using correlation ellipses

#### Variance Inflation Factors (VIFs)

### Part c) Model Output

#### Multiple Regression Model

Multiple Regression Model [conc\_model]: strength ∼ cement + wcr + age

#### Model Output, Adequacy & Appropriateness of Fit

1. **Regression Coefficients**

The **slope** coefficients (cement: 0.06657, wcr: -37.44811, and age: 0.26614) indicate the respective change (increase [+] or decrease [-]) in the concrete strength when each of the predictors increase by 1 unit, but all other predictors remain unchanged.

* The p-values in summary(conc\_model) determines whether the different response-predictor relationships are statistically significant. The p-value are all below 0.05, so we reject the null hypothesis on a 5% significance level and conclude that all the variables (cement, wcr, and age) significantly affect concrete strength. A zero slope coefficient is implausible for all the predictors.

The **intercept** coefficient corresponds to the estimated (theoretical) concrete strength value when all the predictors (cement, wcr, and age) are equal to zero.

* It’s p-value (0.942) is not statistically significant at the 5% level, and an intercept of zero is plausible.
* However, interpreting this is not practically rational but ensures the regression hyperplane fits the data best within the observed predictor values range. It is not meaningful to extrapolate the predictors to zero.

1. **Model Significance**

From the summary (the global F-Statistic), we gather that p-value is very small (4.441e-14) and that the model is highly significant at the 5% level.

1. **Adequacy of Fit [**R2**]**

The R-squared from summary(conc\_model) indicates how much variation in concrete strength is explained by the three predictors as per the regression hyperplane. Here, multiple R2 = 0.6852 (the adjusted R2 = 0.6684), meaning that 69% of the variation in concrete strength is explained by predictors (cement, wcr, and age), while the remaining 31% is due to other factors not included in the model.

1. **Appropriateness of Fit [Model Diagnostics]**
2. **Linearity: E [*Ei*] = 0**

The Tukey-Anscombe residual plot shows that the smoother does not deviate from the x-axis except for a slight () kink for fitted values between 10 and 20 but this deviation can be attributed to randomness. Using the resampling approach by the R function, resplot(), the original red smoother is within what can be generated by random sampling. It is thus imperative to that we accept the linearity hypothesis E [*Ei*] = 0.

Hence, there is no systematic error and the hyperplane is the correct fit.

1. **Homoskedasticity, Var (*Ei*) = *σ2E***

Scale-Location Plot and T-A plot

From the Scale-Location plot, the red smoother is generally horizontal with a gentle kink (between 5 and 17 of the fitted values) which can be considered random. Using the resampling approach, the smoother line is well within the confidence region. We can consider that there is no heteroscedasticity.

1. **No Correlation: Cov (*Ei,Ej*) = 0**

Since the concrete dataset observations are not affected by temporal or spatial variation, the errors can be considered independent and uncorrelated.

1. **Normality: *Ei* ∼ N(0,*σ2E*)**

From the Normal Q-Q Plot, the bulk of the residuals (largely in the central region) are approximately Gaussian distributed. A noticeable deviation (3 outliers) at the upper tail indicates right skewness and departure from normality but because all residuals from the concrete dataset fall within the resampling based confidence region, there is no systematic deviation from the normal distribution. Therefore, the *i.i.d.* assumption holds.

**Summary**: From the R2 value (0.6852), the regression model (hyperplane) is **adequate** because it accounts for a large portion of the total variation in the concrete strength. The model is also **appropriate** because of the good model diagnostics.

### Part c): Variable Selection

Backward Elimination,

Forward Selection,

AIC Stepwise

### Part d): 5-fold Cross Validation & MSPE

Report the mean square prediction error (MSPE)

### Part e): Prediction

Comment on whether this prediction is practically useful.

The model predicts a mean of 11.62271 MPa

The prediction interval spans over 12.6 MPa (from 5.324389 to 17.92103) which reflects high variability in strength for a single batch given the inputs. For structural design, this constitutes a very large uncertainty and the mix may not consistently meet design requirements.

Practically, this result is not fully reliable for decision-making about a specific batch without further testing or improving the model.

## .Question 2

Energy consumption data from 80 office buildings

### Part a): Multicollinearity

### Part b): Model and Predictor Linearity

#### Initial Model Output, Adequacy & Appropriateness of Fit

#### Predictor Linearity

Using partial residual plots

#### Transformed Model, Adequacy & Appropriateness of Fit

### Part c): Variable Selection

Starting from the appropriately transformed model.

Compare results

### Part d): 5-fold cross-validation & MSPE

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: MCQ Answer

**B**. Multicollinearity is present among the predictors.

### Q 3.2: MCQ Answer

D. Cross-validation can help compare models based on predictive accuracy

# Part 2: Analysis of Variance (ANOVA)

Analysis of Variance refers to

## Question 4

## Question 5

## Question 6

Analysis Of Variance theory questions

### Q 6.1

### Q 6.2

### Q.6.3

### Q6.4

# Residual Write-Up

literally may not be meaningful, as real-world conditions rarely involve a speed of exactly zero in this context. its practical importance is limited. how much stopping distance increases per unit increase in speed. A positive slope suggests that higher speeds lead to longer stopping distances. slightly structured residuals. The Normal Q-Q plot suggests that residuals are right-skewed. Log-transformation might therefore be beneficial. The assumption of Gaussian errors is slightly violated by the model due to this moderate non-normality Log-transformation might therefore be beneficial.

indicating that the error variance can be considered constant with fitted values (minor heteroscedasticity). The Tukey-Anscombe plot also seems to indicate that the scatter is not constant for the entire range of speed/fitted values (less scatter for lower values and more scatter for higher values).

Cov (Ei,Ej) = 0

Finally, there must not be any correlation among the errors for different instances, which boils down to the fact that the observations, respectively their errors, do not influence each other, and that there are no latent variables (e.g. time/sequence of the measurements) that do so.

for by the is a fairly adequate fit (high R2), but poor residual plots, verifying that the model is not appropriate

# REFERENCES