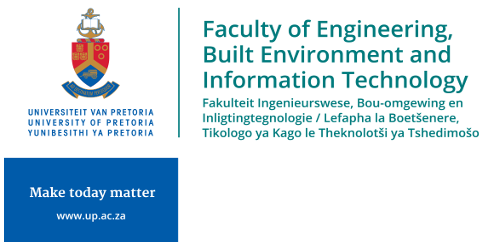
****

**DEPARTMENT OF CIVIL ENGINEERING**

**SHC 798**

**APPLIED STATISTICAL METHODS AND OPTIMISATION**

**Multiple Linear Regression & ANOVA**

**RICHARD LUBEGA**

*Full names*

**25585089**

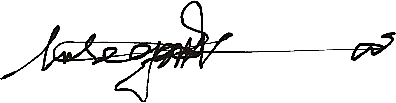
*Student number*

**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

A screenshot of a computer

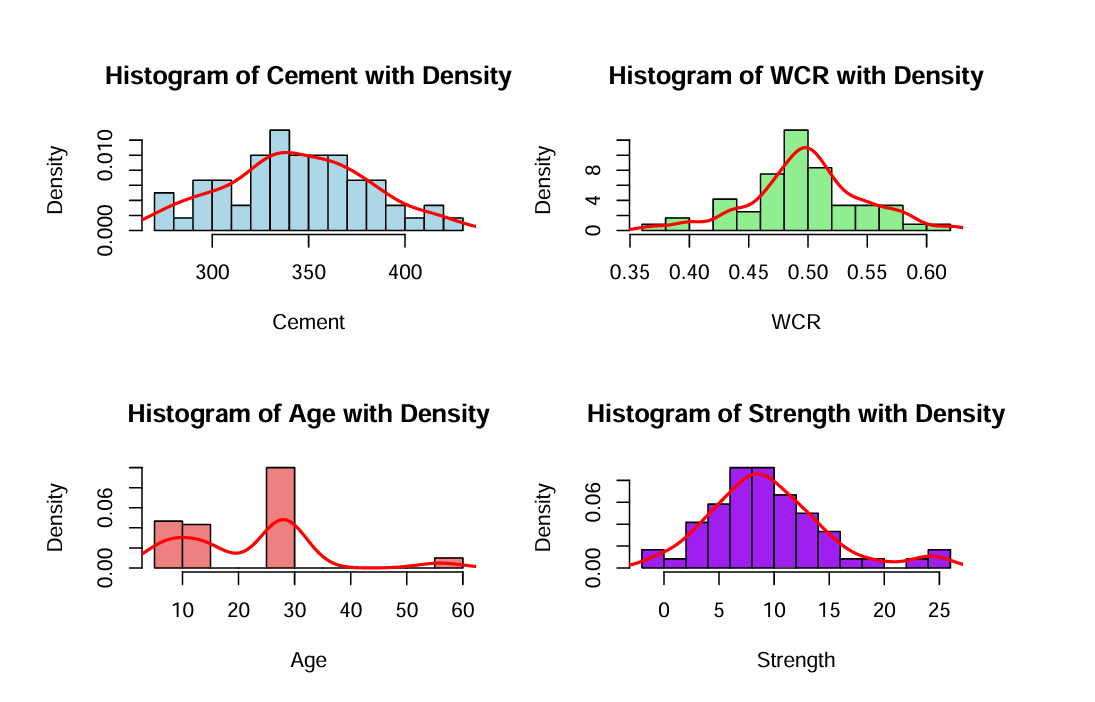
AI-generated content may be incorrect.

### Part a): Data Preparation

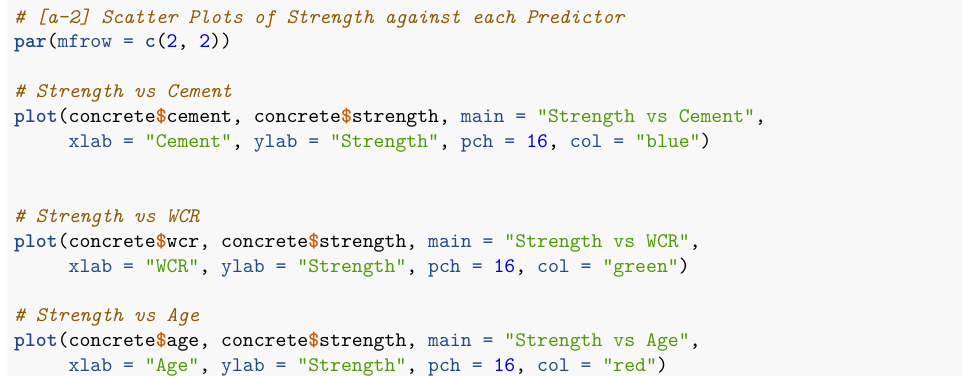
1. Histograms and Marginal Distributions

A screenshot of a computer code

AI-generated content may be incorrect.



1. Scatter Plots



A graph of strength and age

AI-generated content may be incorrect.

**Commenting on the Trend and Need for Variable Transformation**

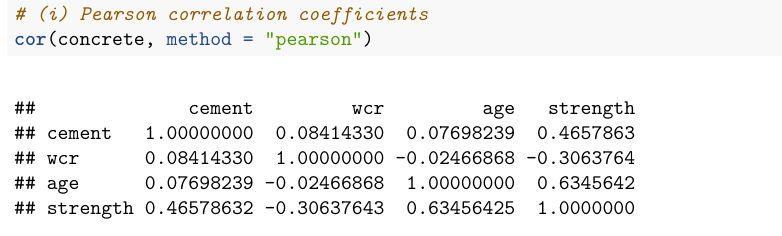
The marginal plots are not skewed and there is no warranted need for variable transformations.

The scatter plot for strength vs age indicates has distinct values (7, 14, 28, 56) which suggests a discrete or categorical nature rather than continuous. The marginal plots for age also show spikes at these specific ages rather than a smooth distribution.

Therefore, age may be as a categorical variable (factor) in regression to account for its discrete levels. Including *interaction terms* (e.g., cement:age, wcr:age) in such a regression model may be also necessary.

### Part b): Multicollinearity

#### Pearson correlation coefficients



#### Ellipse plot to visualise collinearity

A screen shot of a computer

AI-generated content may be incorrect.

#### Variance Inflation Factors (VIFs)

A close-up of a white background

AI-generated content may be incorrect.

**Comment on the findings**

From the above collinearity audit checks (Pearson correlation coefficients and the ellipse plot), the somewhat elongated ellipses, particularly between strength and cement (0.46578632), and strength and age (0.6345642), suggest potential multicollinearity among these predictors.

This indicates that these predictors may be highly correlated with each other and with the response variable, but Since all VIF values are very close to 1 (well below 5), there is no significant multicollinearity among the predictors. This suggests that the predictors are largely independent of each other, which is ideal for a stable regression model.

### Part c) Model Output

#### Multiple Regression Model

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

A screenshot of a computer

AI-generated content may be incorrect.

#### 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) determine 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. **Appropriateness of Fit [Model Diagnostics]**

Residual Plots

1. **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***

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.

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.

**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.

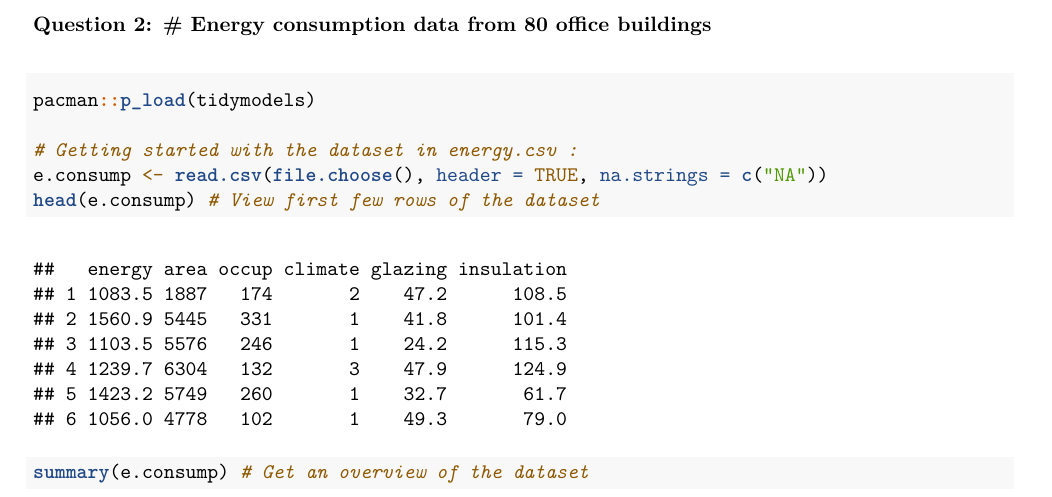
The model predicts a mean of 11.62271 MPa

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



A screenshot of a computer

AI-generated content may be incorrect.

### Part a): Multicollinearity

#### Pearson correlation coefficients

A screenshot of a computer

AI-generated content may be incorrect.

#### Ellipse plot to visualise collinearity

A screenshot of a computer

AI-generated content may be incorrect.

#### Variance Inflation Factors (VIFs)

A close-up of a computer screen

AI-generated content may be incorrect.

**Commenting on Multicollinearity**

In the correlogram (ellipse plot), narrow/elongated ellipses indicate stronger correlation. Energy has elongated ellipses with area (0.5672719) and occupancy (0.71535501), indicating moderate to strong positive correlation. Also, area and occupancy are noticeably correlated with narrow tilted ellipse (0.60076867) which indicates collinearity. Therefore, there is some multicollinearity between area and occupancy, and to a lesser extent between energy and these two variables.

Since all VIF values are very well below 5, there is no significant multicollinearity among the predictors for the model, engy\_model. This suggests that the predictors can be considered independent of each other for this regression model.

### Part b): Model and Predictor Linearity

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

**Multiple Regression Model**

A screenshot of a computer

AI-generated content may be incorrect.

The model is [energy ~ area + occup + climate + glazing + insulation]

1. **Regression Coefficients**

The **slope** coefficients in the engy\_model summary above indicate the respective change (increase [+] or decrease [-]) in energy consumption when each of the predictors increase by 1 unit while all other predictors remain unchanged.

* The p-values determine whether the different response-predictor relationships are statistically significant. Only 3 predictors (area, occup, and insulation) have p-values are all below 0.05 (where we reject the null hypothesis on a 5% significance level) Therefore, these variables significantly affect energy consumption. A zero slope coefficient is plausible for the other predictors (climate and glazing). Hence, they likely do not affect energy consumption

The **intercept** coefficient corresponds to the estimated (theoretical) energy consumption value when all the predictors are equal to zero.

* It’s p-value (8.23e-12) is statistically significant at the 5% level, and an intercept of zero is not plausible.
* Although interpreting this is not practically rational, it 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**

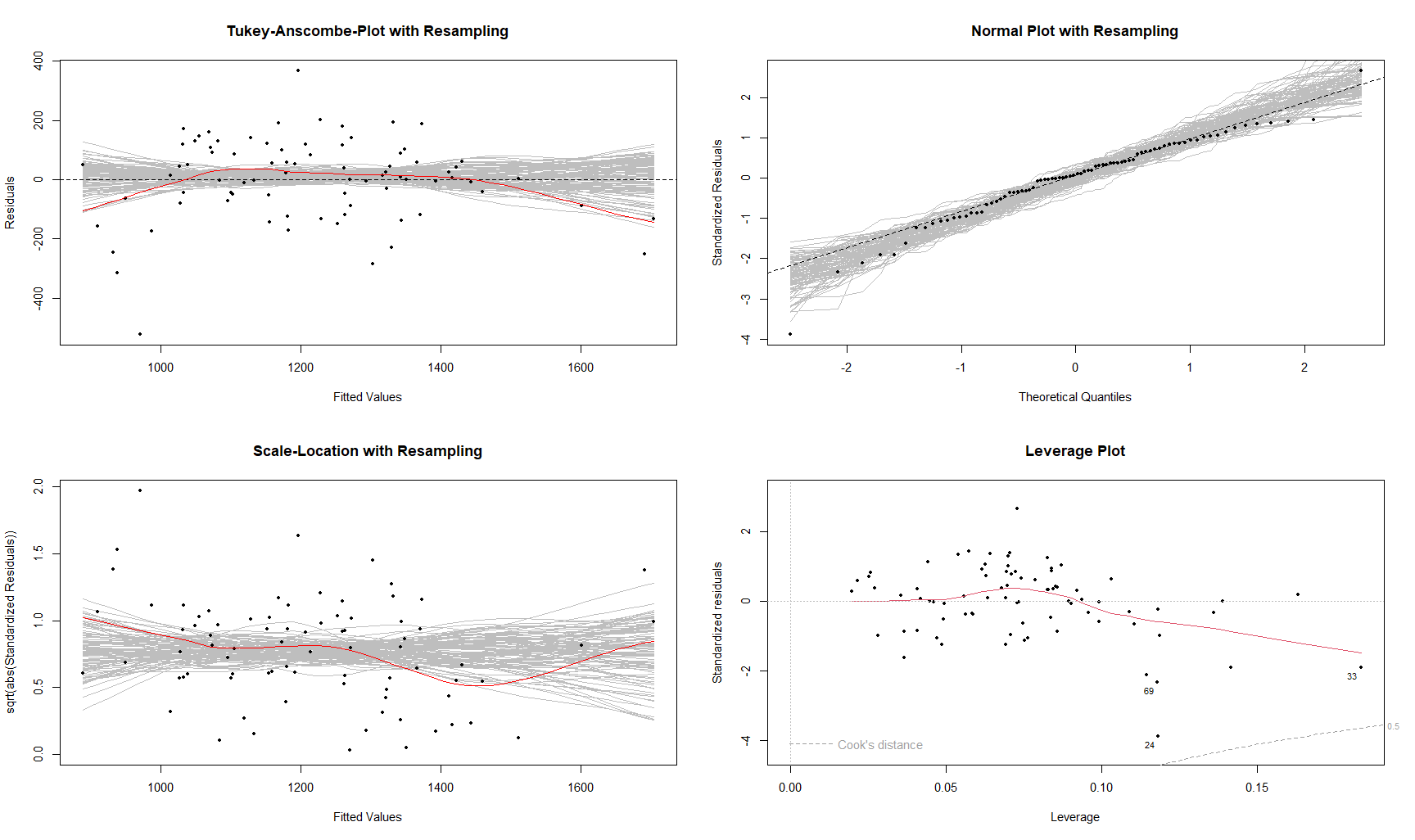
From the summary (the global F-Statistic), we gather that p-value is very small (1.101e-13) and that the model is significant at the 5% level.

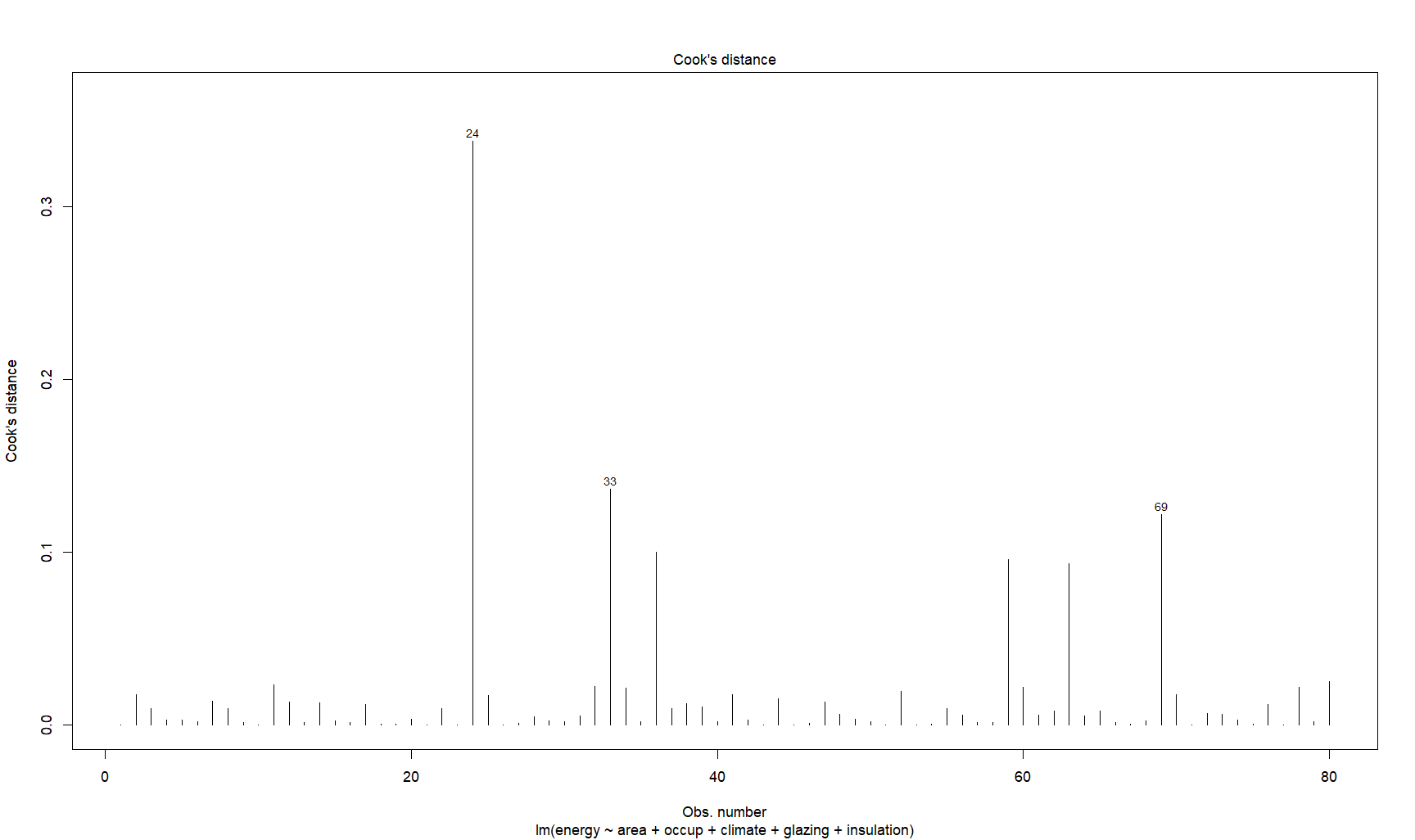
1. **Adequacy of Fit [**R2**]**

The R-squared from summary (engy\_model) indicates how much variation in energy consumption is explained by the five predictors as per the regression hyperplane. Here, multiple R2 = 0.6042, (the adjusted R2 = 0.5774), meaning that 61% of the variation in energy consumption is explained by predictors (area, occup, climate, glazing, and insulation), while the remaining 39% is due to other factors not included in the model.

1. **Appropriateness of Fit [Model Diagnostics]**

Residual Plots





1. **Linearity: E [*Ei*] = 0**

The Tukey-Anscombe residual plot shows that the smoother noticeably deviates from the x-axis at low and high fitted values. From the resampling approach by the R function, resplot(), this deviation may be attributed to randomness because the original red smoother is within what can be generated by random sampling. We accept the linearity assumption.

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

From the Scale-Location plot, the red smoother is generally horizontal and the slight kink (between 1400 and 1600 of the fitted values) can be considered random because the smoother line is well within the resampling confidence region. There is no worrying heteroscedasticity.

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

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

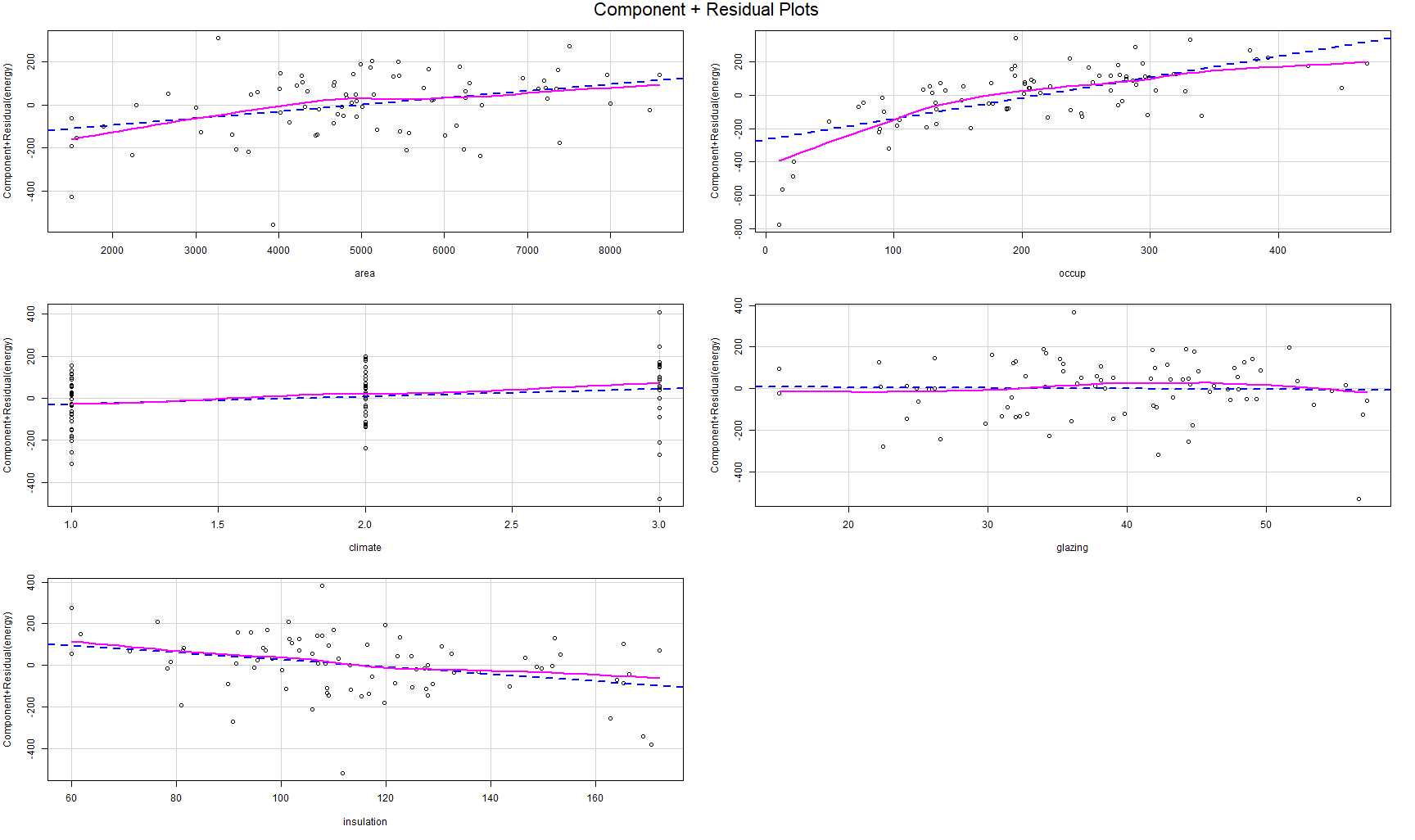
1. **Normality: *Ei* ∼ N(0,*σ2E*)**

From the Normal Q-Q Plot, the bulk of the residuals (largely in the central region) are approximately normally distributed. There are some outliers at both tails which may imply departure from normality. All residuals from this dataset fall within the resampling confidence region, which means that deviations are random. The normality*.* assumption holds.

**Summary**: The model is also **appropriate** because of its associated residual plots are acceptable. The R2 value (0.6042) implies that the regression model (hyperplane) is **adequate** because it accounts for a large portion of the total variation in the energy consumption.

#### Predictor Linearity

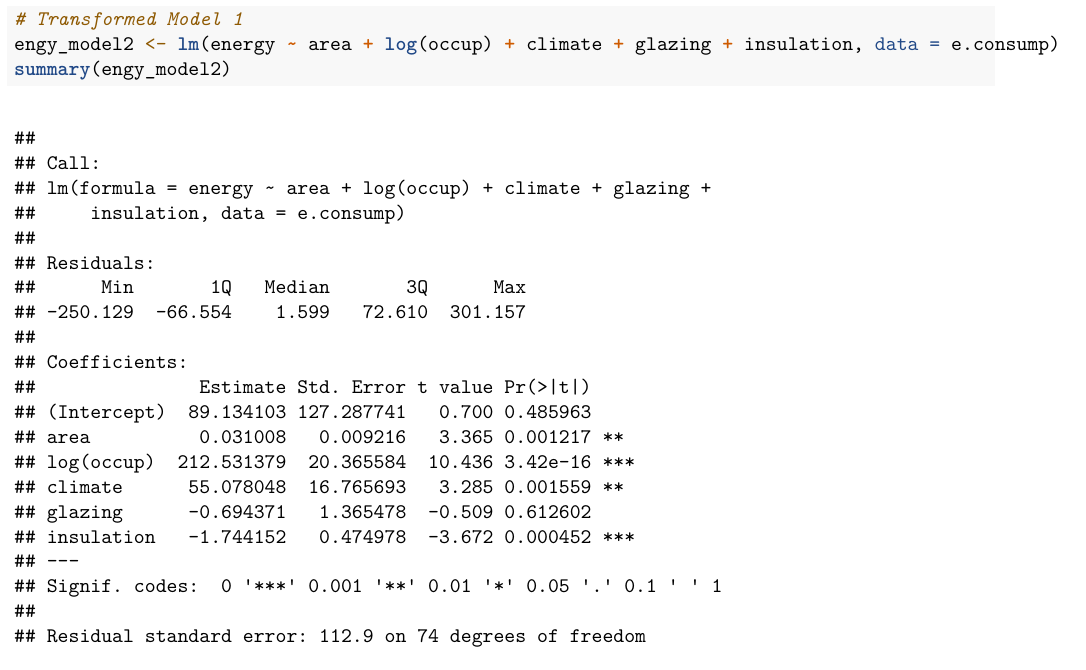
The partial residual plots are shown for the initial/original model (engy\_model).



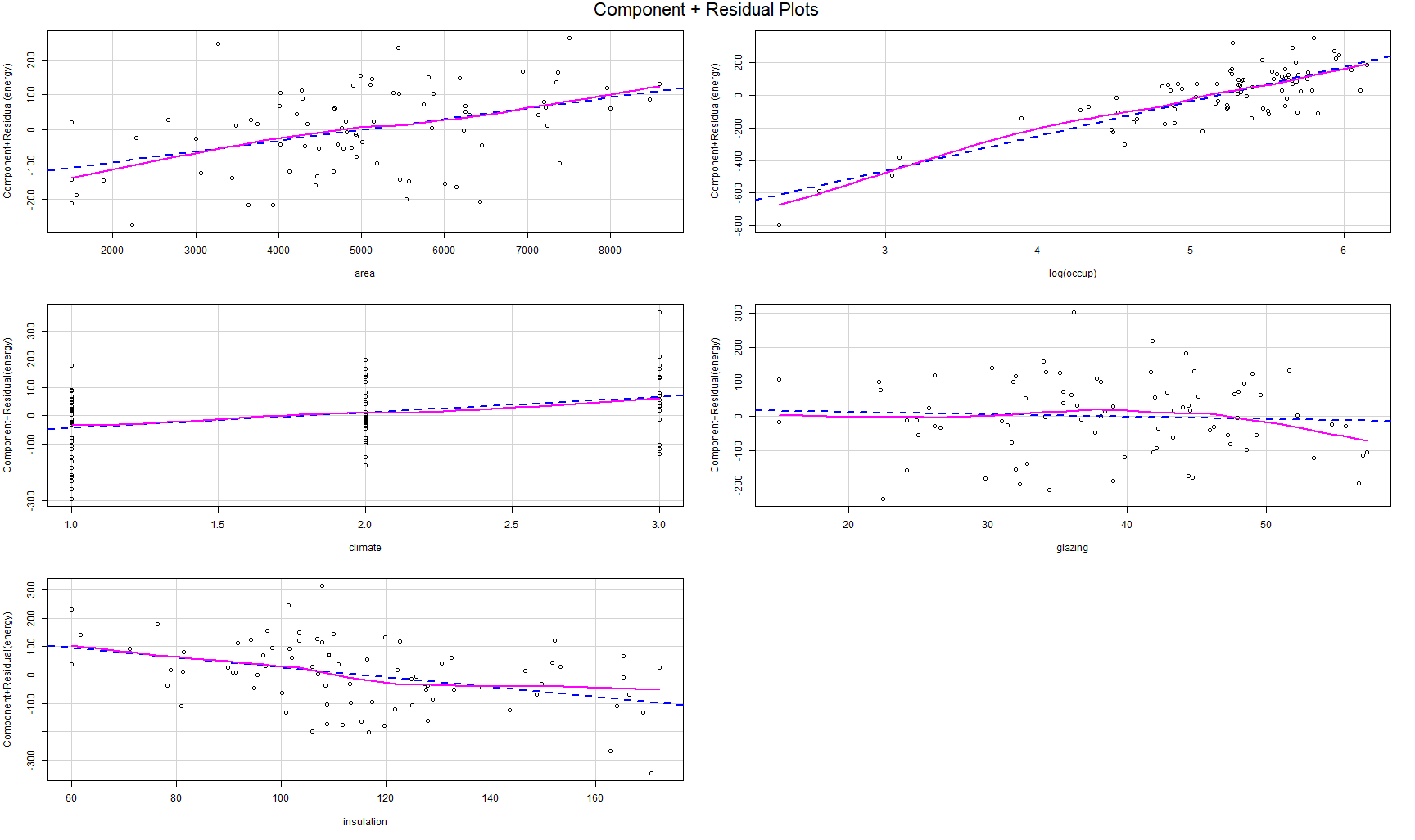
From the partial plots of the initial/original (engy\_model) above, predictors, the variables area and occupancy clearly deviate from the blue dotted line which indicates non-linearity.

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

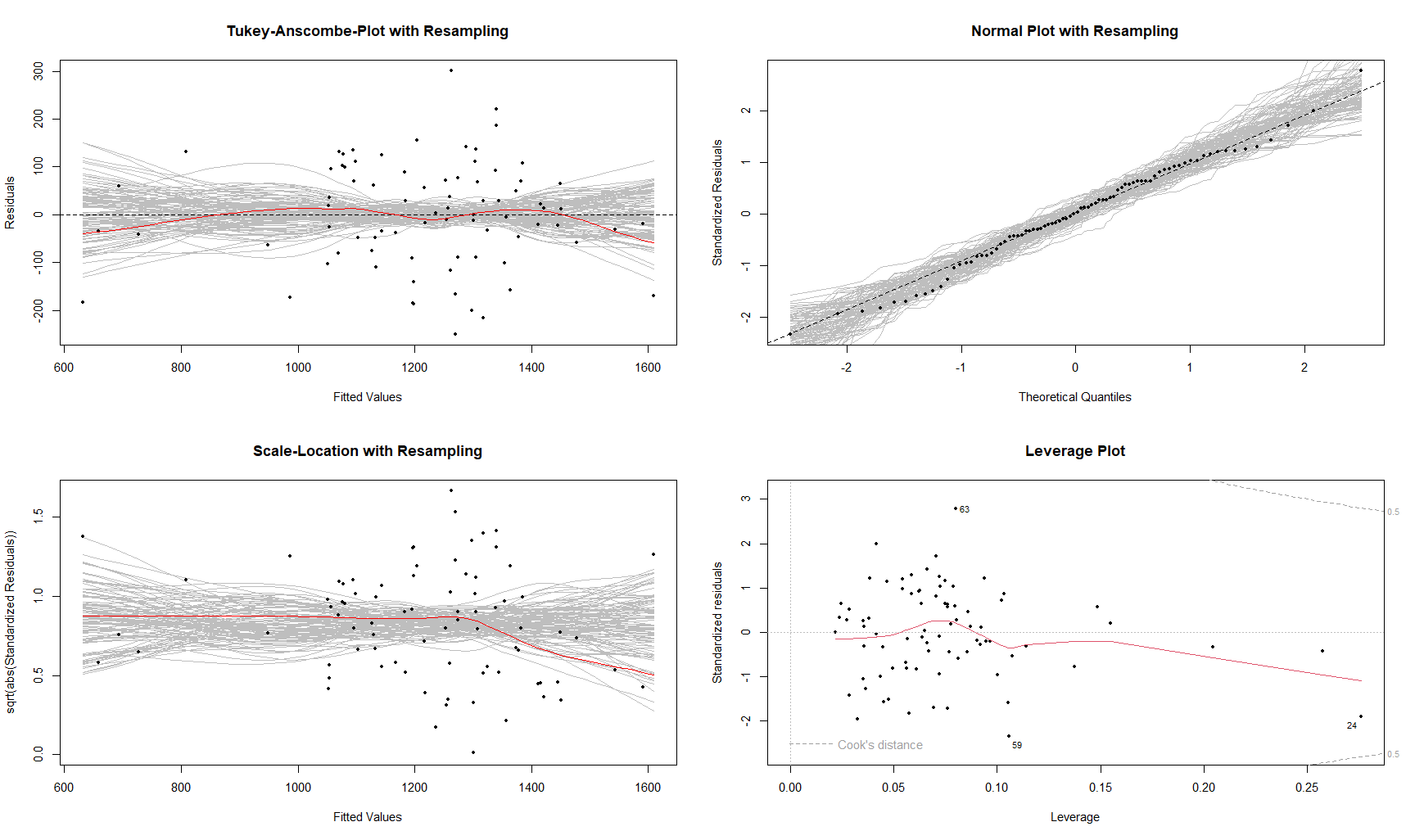
**Transformed Model 1** (engy\_model2). Here the occup is log-transformed



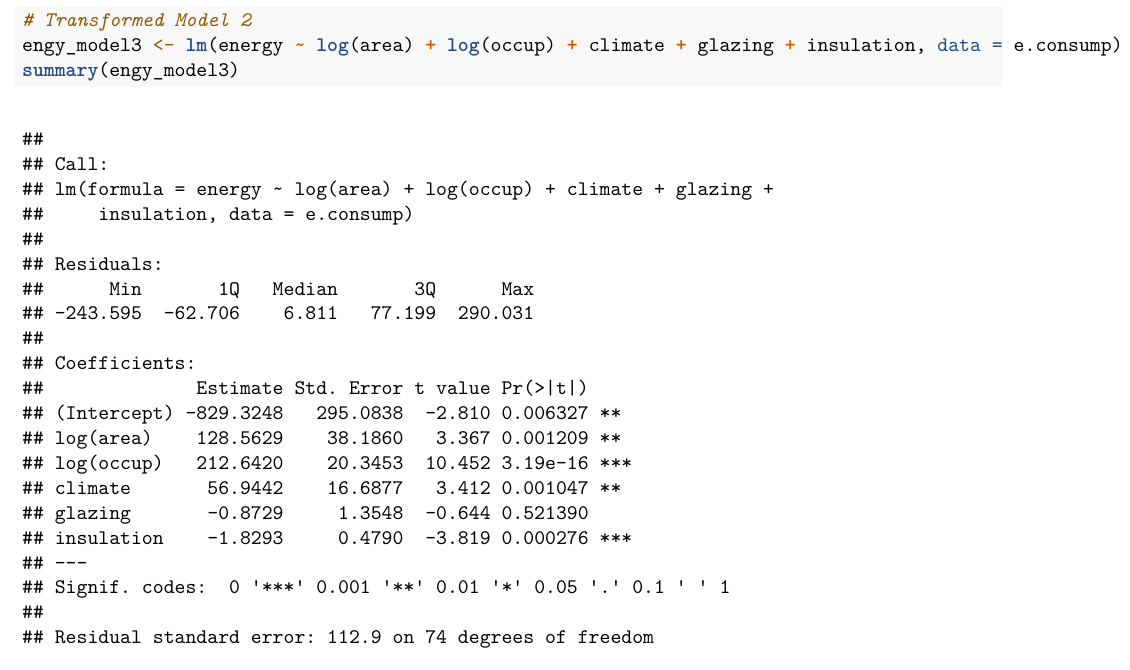
Partial Plots for engy\_model2



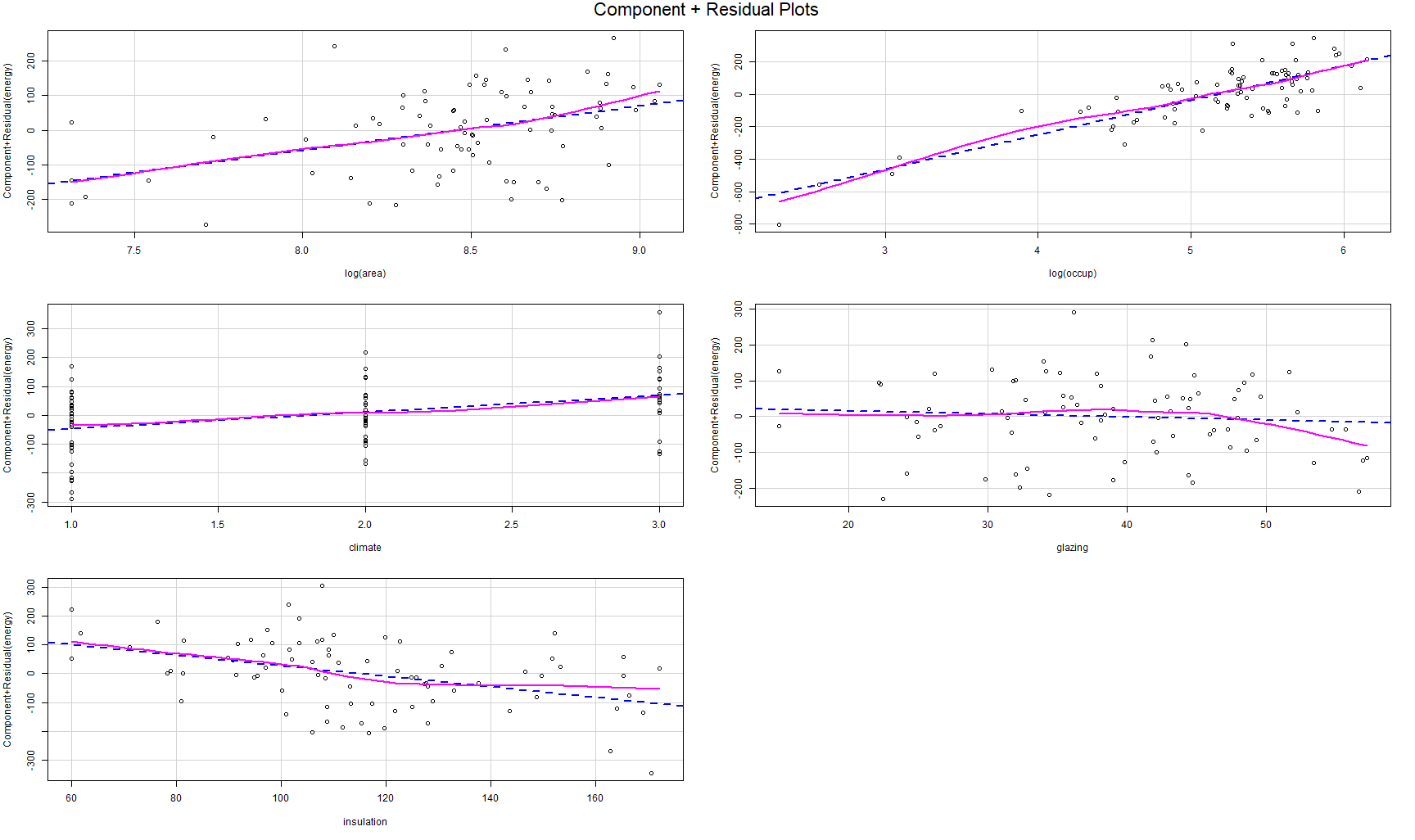
Residual Plots for engy\_model2



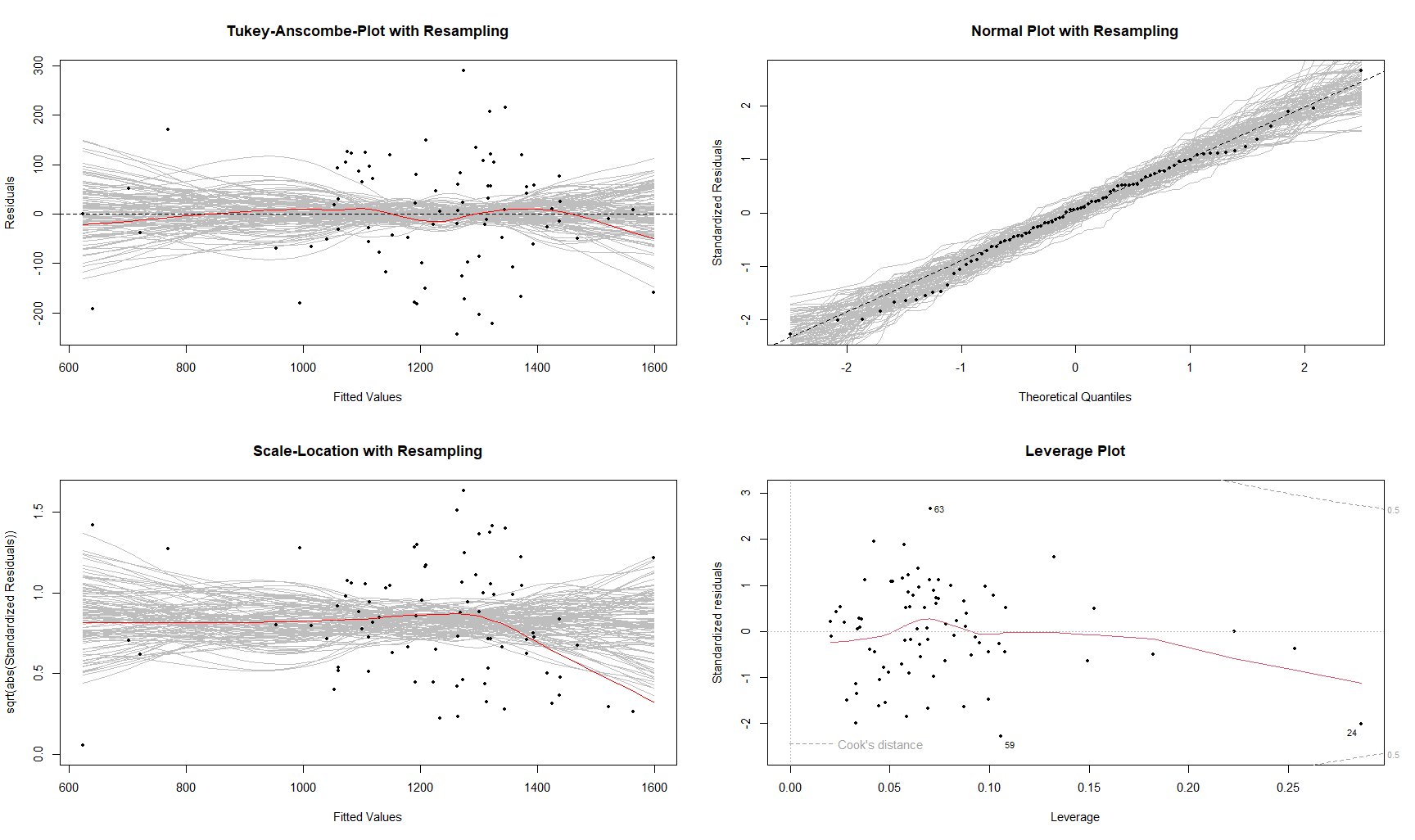
**Transformed Model 2** (engy\_model3). In this model, area and occup are log-transformed



Partial Plots for engy\_model3



Residual Plots for engy\_model3



**Commenting on Model outputs, adequacy of fit and appropriateness of fit**.

In the first transformed model (engy\_model2), the linearity of both variables are seen to improve. Also, the model diagnostics (appropriateness of fit) are much better for this transformed model. From the Adjusted R2, this model also fits the data better (0.7371) than the original/initial model ( 0.5774).

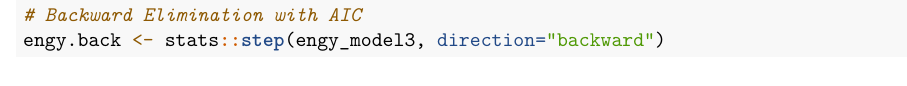
In the second transformed model (engy\_model3), the variable linearity, residual plots (appropriateness of fit) and model fit are better than both the original and the first transformed model (engy\_model2)

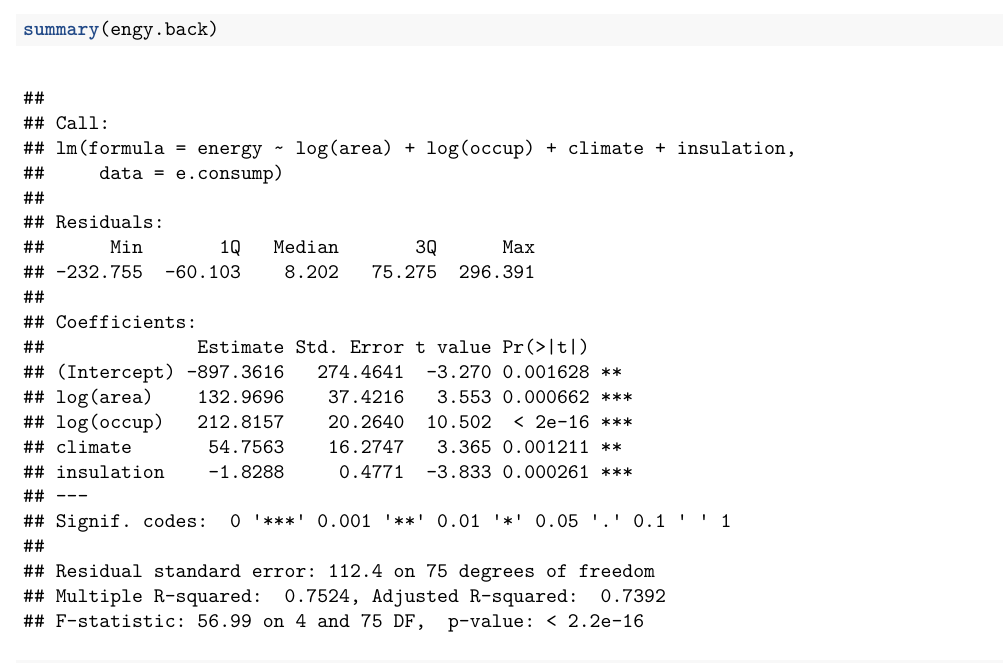
Therefore, this model (engy\_model3), is taken as the most appropriate in this case.

### Part c): Variable Selection

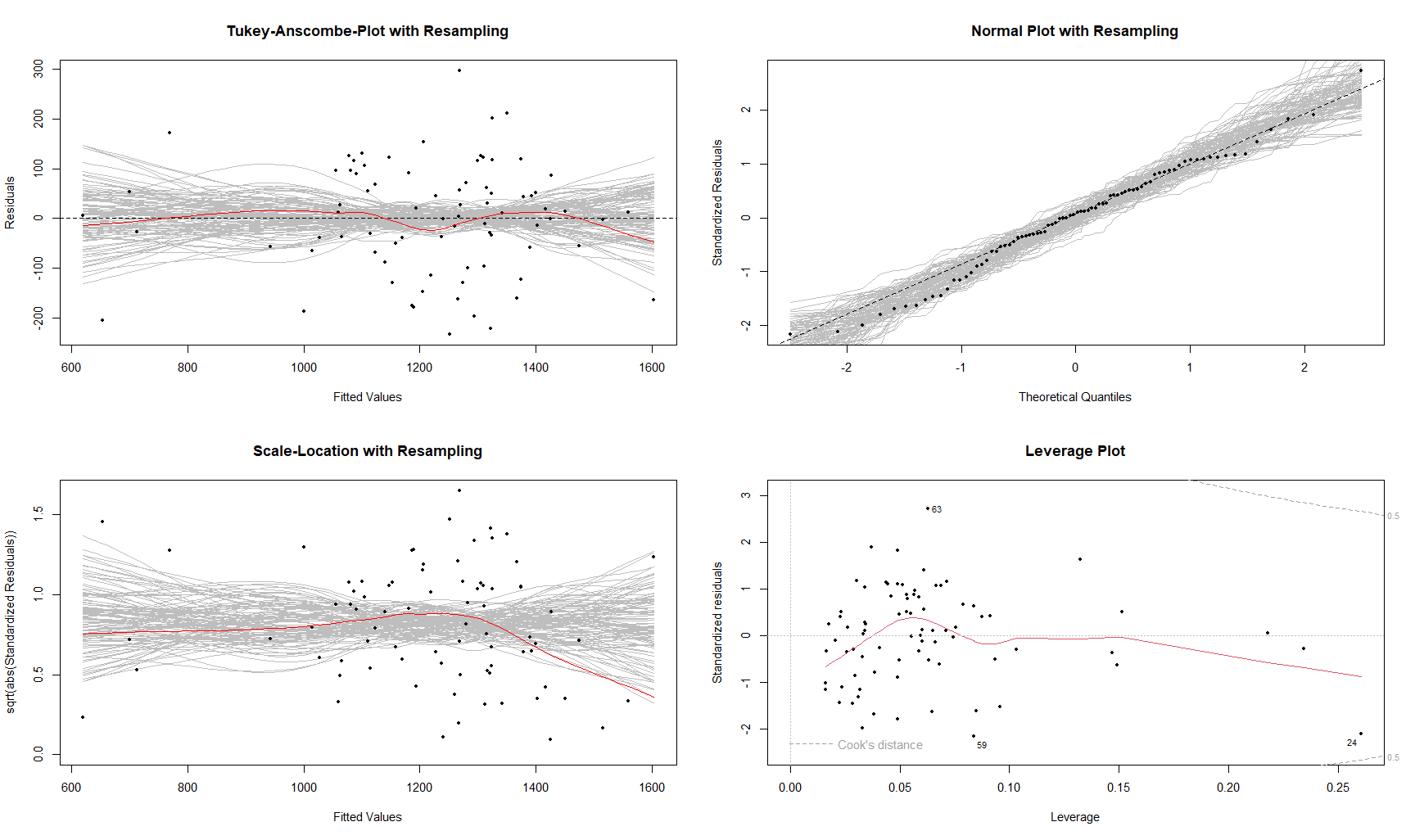
Starting from the appropriately transformed model (engy\_model3).

1. **Backward Elimination Model** (engy.back)





Model Diagnostics (Residual Plots) for engy.back

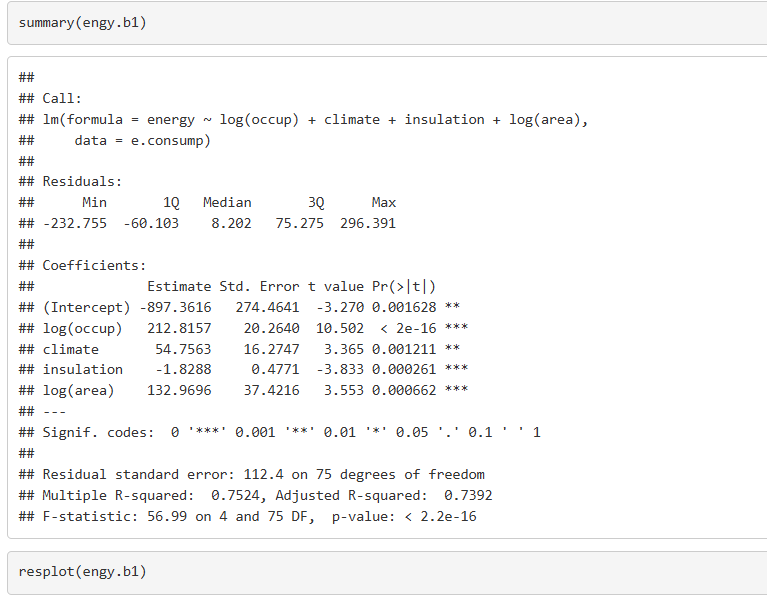


1. **AIC Stepwise Models** [engy.b1, engy.b2, and engy.b3]

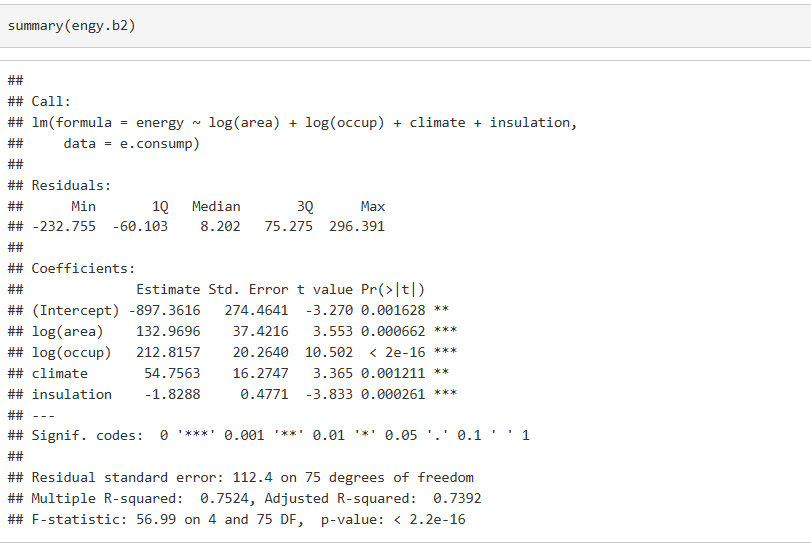
A close-up of a math equation

AI-generated content may be incorrect.

Model engy.b1



Model engy.b2



Model engy.b3

A screenshot of a computer

AI-generated content may be incorrect.

**Comparing results**

In all the reduced models from applying variable selection (i.e., engy.back, engy.b1, engy.b2 and engy.b3), the variable, glazing, was dropped.

There are no major improvements in residual plots for all the models (here, only plots for the model engy.back are shown). Also, no noticeable changes (improvements) on the remaining predictor significance or model fit as compared to the full transformed model (engy\_model3).

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

Compute MSPE for both the full and the reduced model. Which performs better for prediction?

**The 5-fold cross-validation loop code**

A screenshot of a computer program

AI-generated content may be incorrect.

**MSPE values** for both the full and reduced models

* Full Model



* Reduced Model



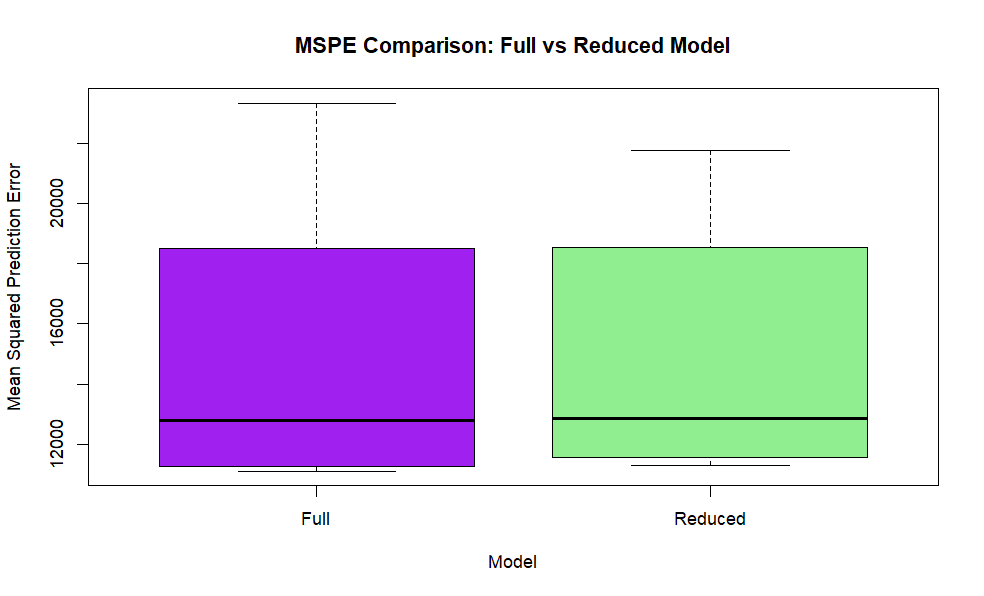
* Comparing change in MSPEs (Full to Reduced)



Visualising MSPEs with Box Plots

A screen shot of a computer code

AI-generated content may be incorrect.



**Comparing the models**

From the cross-validation exercise, The MSPE for the *reduced* model is less (-1.275735%) than the *full* model. This implies that the variable, glazing, can be said to reduce the predictive power in model.

Thus in this case, the reduced model is preferable for prediction purposes.

## 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