

DATA410 Project Report

Dylan Tan and Irfan Iskandar

2024-04-10

Predicting core body temperature using infrared thermography (IRT)

1. Background Information

In recent years, the medical field has seen a significant increase in the use of non-invasive diagnostic techniques, which are techniques that do not require the introduction of instruments into the body, to get real-time estimates of a patient's core body temperature, with infrared thermography (IRT) becoming an important tool for physiological monitoring and disease detection. This is particularly useful for clinical settings and during infectious disease epidemics. For example, in the year 2020 during the COVID-19 pandemic, the Kuala Lumpur International Airport (KLIA) utilized thermal scanners to detect passenger's body temperatures to determine if they were potential carriers of the COVID-19 disease. Only passengers with a core body temperature of below 37 degrees Celsius were allowed to move into the airport's check-in area (Loo, 2020).

The non-invasive nature of IRT enables its extensive utilization in various medical applications, as abnormal body temperature serves as a natural indicator of illnesses. Biomedical research has showcased the efficiency of IRT in diverse diagnostic endeavors, including the detection of breast cancer, diabetes neuropathy, peripheral vascular disorders, gynecological issues, kidney transplantation, dermatological conditions, cardiac abnormalities, neonatal physiology, fever screening, and brain imaging (Kylili et al., 2014).

IRT's capability to visualize and quantify changes in surface temperatures has led to its application in monitoring a wide range of medical conditions, such as severe acute respiratory syndrome, Ebola virus disease, or even the coronavirus disease in 2019. A crucial aspect of using IRT lies in the accurate interpretation of thermal images, particularly in correlating external thermal readings with core body temperatures.

Our project aims to close this gap by developing a predictive model that can estimate oral temperature—a common measuring technique of core body temperature—using thermal imaging data from IRTs and environmental factors such as ambient temperature, relative humidity, and the distance between the subjects and the IRTs. This model could potentially enhance the clinical use of IRT by providing non-invasive measurement methods, offering a significant advantage in both routine health assessments and the early detection of health problems, as well as the detection of carriers of infectious diseases.

We hypothesize that by using environmental factors and IRTs, we can predict oral temperature.

2. Data Description

The Infrared Thermography Temperature Dataset is designed for regression analysis to predict oral temperature based on environmental conditions and thermal image readings. It includes temperatures from various locations in infrared images of patients, complemented with oral temperatures measured for each individual. The dataset consists of 33 features, covering gender, age, ethnicity, ambient conditions (temperature and humidity), distance from the infrared thermography (IRT) device, and various other temperature readings from thermal images. This rich collection of data points serves as a critical tool for understanding the environmental and physiological factors influencing oral temperature.

Variables Table

Name	Role	Type	Description
SubjectID	ID	Categorical	Subject ID
aveOralF	Target	Continuous	Oral temperature measured in fast mode
aveOralM	Target	Continuous	Oral temperature measured in monitor mode
Gender	Feature	Categorical	Male or Female
Age	Feature	Categorical	Age ranges in categories
Ethnicity	Feature	Categorical	American Indian or Alaska Native, Asian, Black or African American, Hispanic/Latino, Multiracial, Native Hawaiian or other Pacific Islander, White
T_atm	Feature	Continuous	Ambient temperature
Humidity	Feature	Continuous	Relative humidity
Distance	Feature	Continuous	Distance between the subjects and the IRTs
T_offset1	Feature	Continuous	Temperature difference between the set and measured blackbody temperature
Max1R13_1	Feature	Continuous	Max value of a circle with diameter of 13 pixels from the right canthus point to the face centerline
Max1L13_1	Feature	Continuous	Max value of a circle with diameter of 13 pixels from the left canthus point to the face centerline
aveAllR13_1	Feature	Continuous	Average value of a circle with diameter of 13 pixels from the right canthus point to the face centerline
aveAllL13_1	Feature	Continuous	Average value of a circle with diameter of 13 pixels from the left canthus point to the face centerline
T_RC1	Feature	Continuous	Average temperature of the highest four pixels in a square of 24x24 pixels around the right canthus
T_RC_Dry1	Feature	Continuous	Average temperature of the highest four pixels in the right canthus dry area
T_RC_Wet1	Feature	Continuous	Average temperature of the highest four pixels in the right canthus wet area
T_RC_Max1	Feature	Continuous	Max value of a square of 24x24 pixels around the right canthus
T_LC1	Feature	Continuous	Average temperature of the highest four pixels in a square of 24x24 pixels around the left canthus
T_LC_Dry1	Feature	Continuous	Average temperature of the highest four pixels in the left canthus dry area
T_LC_Wet1	Feature	Continuous	Average temperature of the highest four pixels in the left canthus wet area
T_LC_Max1	Feature	Continuous	Max value of a circle with diameter of 13 pixels from the left canthus point to the face centerline
RCC1	Feature	Continuous	Average value of a square of 3x3 pixels centered at the right canthus point

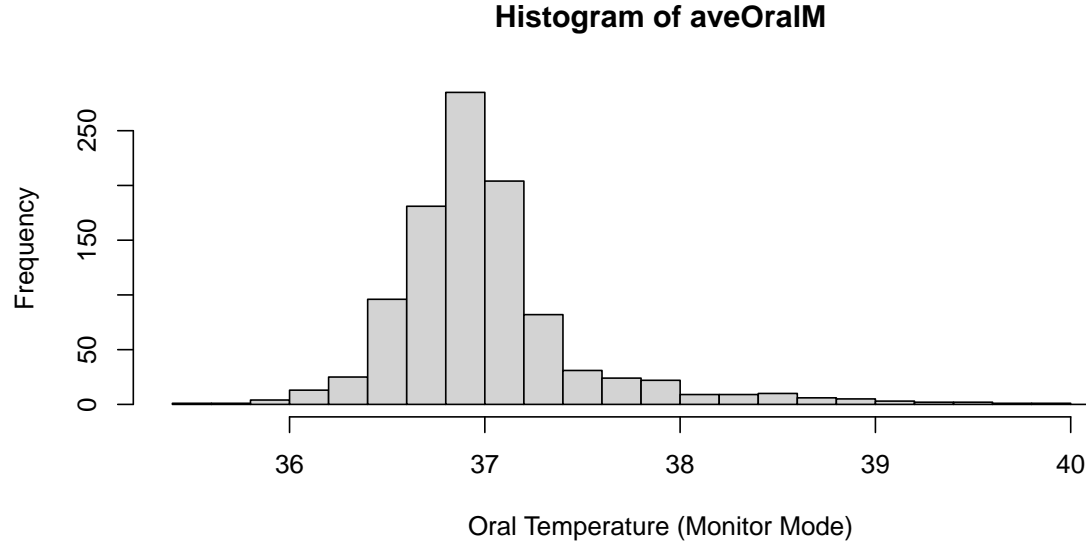
Name	Role	Type	Description
LCC1	Feature	Continuous	Average value of a square of 3x3 pixels centered at the left canthus point
canthiMax1	Feature	Continuous	Max value in the extended canthi area
canthi4Max1	Feature	Continuous	Average temperature of the highest four pixels within the extended canthus area
T_FHCC1	Feature	Continuous	Average value in the center point of forehead
T_FHRC1	Feature	Continuous	Average value in the right point of the forehead
T_FHLC1	Feature	Continuous	Average value in the left point of the forehead
T_FHBC1	Feature	Continuous	Average value in the bottom point of the forehead
T_FHTC1	Feature	Continuous	Average value in the top point of the forehead
T_FH_Max1	Feature	Continuous	Maximum temperature within the extended forehead area
T_FHC_Max1	Feature	Continuous	Max value in the center point of forehead
T_Max1	Feature	Continuous	Maximum temperature within the whole face region
T_OR1	Feature	Continuous	Average temperature of the highest four pixels within the mouth region
T_OR_Max1	Feature	Continuous	Maximum temperature within the mouth region

3. Regression Analysis

3.1: Data Preparation and EDA

After importing the dataset, we merged the target variables and features into one dataframe. The dataset consists of three categorical variables: Age, gender, and ethnicity. In order to run a regression model, these variables needed to be coded into dummy variables for interpretability. We also categorized Age and stored it as levels. For the ethnicity_dummy variable, we decided to have binary inputs consisting of white and non-white ethnicity. This decision was made by assessing the distribution of ethnicity in our sample, where we found that white ethnicity was the majority. Additionally, we combined two age groups (i.e., 21-25, 26-30) to prevent overlap with another age group (i.e., 21-30).

Since we created dummy variables for gender and ethnicity, we excluded the original non-factored versions of the variables. Additionally, there were two missing values in the dataset, which we simply remove as the amount of missing data is minimal. The 2 missing values were from the distance column.



Distribution of aveOralM

After plotting the histogram of aveOralM, we found that the distribution of our target variable is right skewed. We initially tried to perform a log transformation but found that the distribution did not change much. Therefore, we decided to assess the dataset with two different models, first being a Gaussian family distribution and the second being a Gamma family distribution with a log link function.

3.2: Initial Variable Selection

In this project, we investigated a few different methods of variable selection. Before delving into all the different methods used, we first decided to explore the variables and found that the descriptions of certain variables are highly correlated to each other. For example: RCC1(Average value of a square of 3x3 pixels centered at the right canthus point), LCC1(Average value of a square of 3x3 pixels centered at the left canthus point) and canthiMax1 (Max value in the extended canthi area) are all similar in nature. Since they are similar in nature, we decided to group them together and select the most general variable as the representative for their group for our model fitting. We used the `corr()` function to find the correlation between the members of the grouped variables, and if the grouped predictor variables were correlated with each other, we selected the one that was most correlated to the target variable to be the representative for that group. From 33 predictor variables, we decided to pick 15 variables for our regression model.

From our initial findings on our target variable, we made a Gaussian family linear model and a Gamma family linear model with a log link function, both using all the 15 chosen predictor variables. We found that the Gamma model had a lower AIC of 216.83 compared to 195.64 from the Gaussian model, which might suggest that it is the better model for now compared to the two. However, both the null deviance and the residual deviance for the Gamma model were low, indicating potential modeling errors. Since the AIC between the two models did not have much of a difference, we decided to investigate further.

Removal of variables with high VIF values.

To prevent the issue of multicollinearity, we decided to remove the variables with high VIFs. Since the variables with high VIFs were the same for both models, the reduced final model was identical. After two iterations of model reduction, we ended up with age, T_atm, Humidity, Distance, T_offset1, Max1R13_1, Max1L13_1, T_FH_Max1, T_FHC_Max1, T_OR_Max1, Gender_dummy, Ethnicity_dummy as our predictor variables for the final model. Table 3.2.1 shows the VIF values of our final model using the VIF variable reduction method.

Table 3.2.1 : VIF table of full model

Table 2: **Variance Inflation Factors (VIFs) of Final Model**

	Predictors	VIF.GVIF	VIF.Df	VIF.GVIF..1..2.Df..
Age	Age	1.124709	5	1.011822
T_atm	T_atm	1.756811	1	1.325448
Humidity	Humidity	1.102517	1	1.050008
Distance	Distance	1.104435	1	1.050921
T_offset1	T_offset1	1.754810	1	1.324693
Max1R13_1	Max1R13_1	6.310592	1	2.512089
Max1L13_1	Max1L13_1	6.033408	1	2.456300
T_FH_Max1	T_FH_Max1	6.121836	1	2.474234
T_FHC_Max1	T_FHC_Max1	6.344559	1	2.518841
T_OR_Max1	T_OR_Max1	2.416082	1	1.554375
Gender_dummy	Gender_dummy	1.106909	1	1.052097
Ethnicity_dummy	Ethnicity_dummy	1.133668	1	1.064738

3.3: Forward Selection and Backward Elimination Variable Selection

Since our previous approach of removing variables with high VIFs did not improve our AIC, we decided to employ the Forward and Backward Elimination using AIC algorithms for model selection to test if we can improve our model.

3.3.1 Linear Model

We found that both methods led to the same set of variables, and reduced the total number of variables from 15 to 8.

Chosen predictor variables for Best model using Forward Selection and Backwards Elimination

$$aveOralM \sim T_OR_Max1 + Max1R13_1 + T_atm + T_FH_Max1 + Max1L13_1 + Ethnicity_dummy + Gender_dummy + T_FHC_Max1$$

Table 3.3.2: Final step and variables selected using Forward Substitution Backward Elimination

Forwards Selection Results

Step: AIC = 250.7

Step	Df	Deviance	AIC
	0	0.054498	250.70
+ T_offset1	1	0.054402	250.93
+ Humidity	1	0.054468	252.15
+ Distance	1	0.054479	252.35
+ Age	5	0.054342	257.82

Backward Elimination Results

Step: AIC = 272.74

Step	Df	Deviance	AIC
		76.344	272.74
- T_FHC_Max1	1	76.844	277.37
- Gender_dummy	1	76.892	278.01
- Ethnicity_dummy	1	77.563	286.85
- Max1R13_1	1	78.164	294.69
- Max1L13_1	1	78.342	297.01
- T_FH_Max1	1	79.208	308.19
- T_atm	1	82.115	344.84
- T_OR_Max1	1	96.768	511.84

The AIC based on the variable selection we did is now 272.74, which is a slight improvement from our initial full model with no variable selection done to it, which had an AIC of 282.9503. However, a new problem arises. We now have a few variables with higher VIFs than our set threshold of 5. They are Max1R13_1, Max1L13_1, T_FH_Max1 and T_FHC_Max1. However, since our VIF values are not much higher than 10, we can still consider this model as acceptable.

Although the problem of multicollinearity might be potentially be an issue, this is an improvement from our initial model.

3.3.2 Forward Selection and Backward Elimination Method with Gamma Model

Forward Selection using AIC for Gamma Model

Step: AIC=250.7

Step	Df	Deviance	AIC
		0.054498	250.70
+ T_offset1	1	0.054402	250.93
+ Humidity	1	0.054468	252.15
+ Distance	1	0.054479	252.35
+ Age	5	0.054342	257.82

Backward Elimination using AIC for Gamma Model

Step: AIC=250.7

Step	Df	Deviance	AIC
		0.054498	250.70
- T_FHC_Max1	1	0.054834	254.90
- Gender_dummy	1	0.054917	256.43
- Ethnicity_dummy	1	0.055356	264.55
- Max1R13_1	1	0.055781	272.39
- Max1L13_1	1	0.055914	274.85
- T_FH_Max1	1	0.056476	285.22
- T_atm	1	0.058564	323.78
- T_OR_Max1	1	0.069161	519.43

Both Forward Selection and Backwards Elimination showed the same model for the Gamma model, with the predictor variables being T_atm + Max1R13_1 + Max1L13_1 + T_FH_Max1 + T_FHC_Max1 + T_OR_Max1 + Gender_dummy + Ethnicity_dummy

The AIC based on both Forward Selection and Backwards Elimination showed the same model for the Gamma model with the same AIC of 250.7. This is an improvement from our initial full model which had an AIC of 261.01. However, a new problem arises. We now have a few variables with higher VIFs than our set threshold of 5. They are T_FHC_Max1, Gender_dummy, Ethnicity_dummy, Max1R13_1, Max1L13_1, T_FH_Max1, T_atm and T_OR_Max1. However, since our VIF values are not much higher than 5, we can still consider this model as acceptable.

3.4 Best Subset Model Method

Using the Best Subset Regression method for variable selection, there were 4 suggested variables, Max1R13_1, Max1L13_1, T_FH_Max1 and Gender_dummy. For the Gaussian Linear Model, we found an AIC of 594.97. For the Gamma Model, we found an AIC of 572.64.

Both of these models had noticeably high increase from the base model, which indicates that this model should not be used.

3.5 Lasso and Ridge

Linear Model LASSO

For lasso we tried with two approaches one being lambda min and the other with the minimization of standard error values. The AIC for lambda min is 277.37 which is a slight decrease from the original model and all the variables in the model this time appears significant with 7 out of the 15 variables being retained. For lambda se1, we found an AIC of 296.55, which is an increase from the base model. However, all the variables are significant and the VIFS are fairly low with only two slightly exceeding 5.

In our case, with lambda min, the number of features was reduced from 12 to 7, while lambda se1 reduced it to 5. Upon checking with the VIFs, the model with fewer predictors generally exhibited lower VIFs. However, choosing the 7-variable model wouldn't be a bad choice because the VIFs are still tolerable. Ultimately, this decision would depend on our objectives. If we aim for a more robust model, we could opt for lambda min. Hence, if we prefer a simpler model that is easier to interpret and more likely to be stable due to the minimization of SE, then lambda se1 would be more suitable.

Linear Model Ridge

With the ridge method we selected features that had a coefficient of more than 0.1 which gave us the predictor variables Max1R13_1, Max1L13_1, T_FH_Max1, T_OR_Max1. These were based on the coefficients we have obtained by setting the $\alpha = 0$ and using the both min and se lambda approach. This resulted in a four variable model.

THE AIC we got from this lamda min approach was 377.13 which was a small jump from the base model . The VIF was acceptable with no variables exceeding the threshold of 10.

Gamma Model

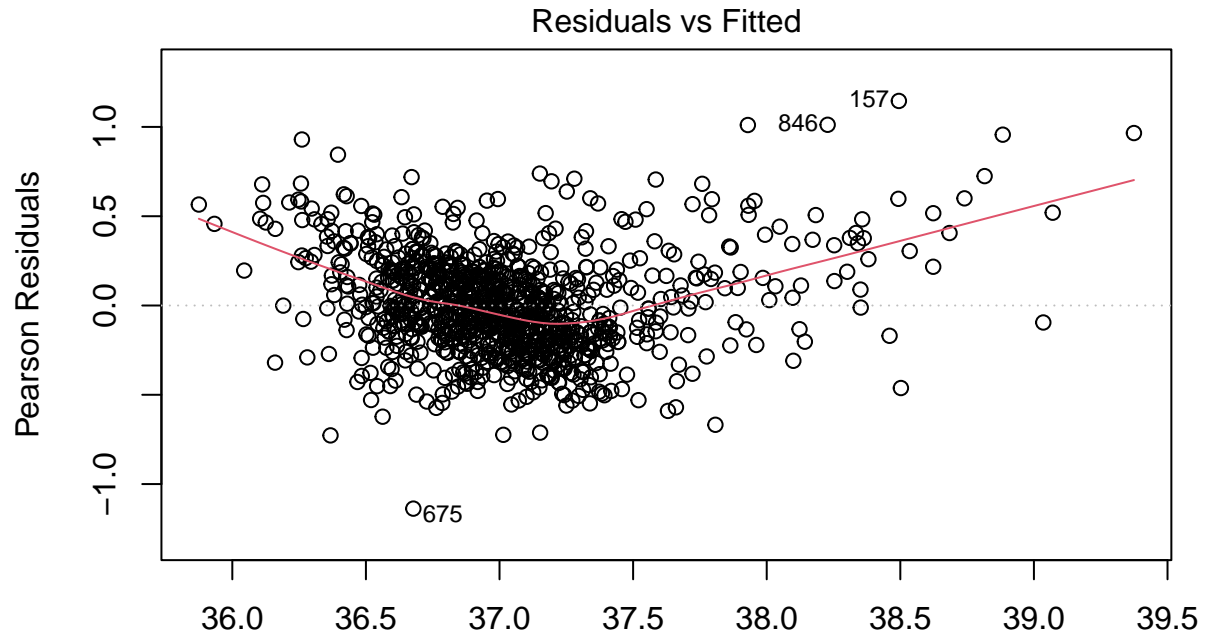
For LASSO for the Gamma family model, we tried with two approaches, the first being lambda min and the second being the minimization of standard error values. However, there was only 1 variable that was chosen from the lambda min approach and the se1 approach gave an intercept only model.

With the ridge method, the predictor variables that had a coefficient of more than 0.1 were selected. These variables were Max1R13_1, Max1L13_1, T_FH_Max1 and T_OR_Max1. these were based on coefficients we've obtained by setting $\alpha = 0$ and using both min and se lambda approaches. This also resulted in a four variable model.

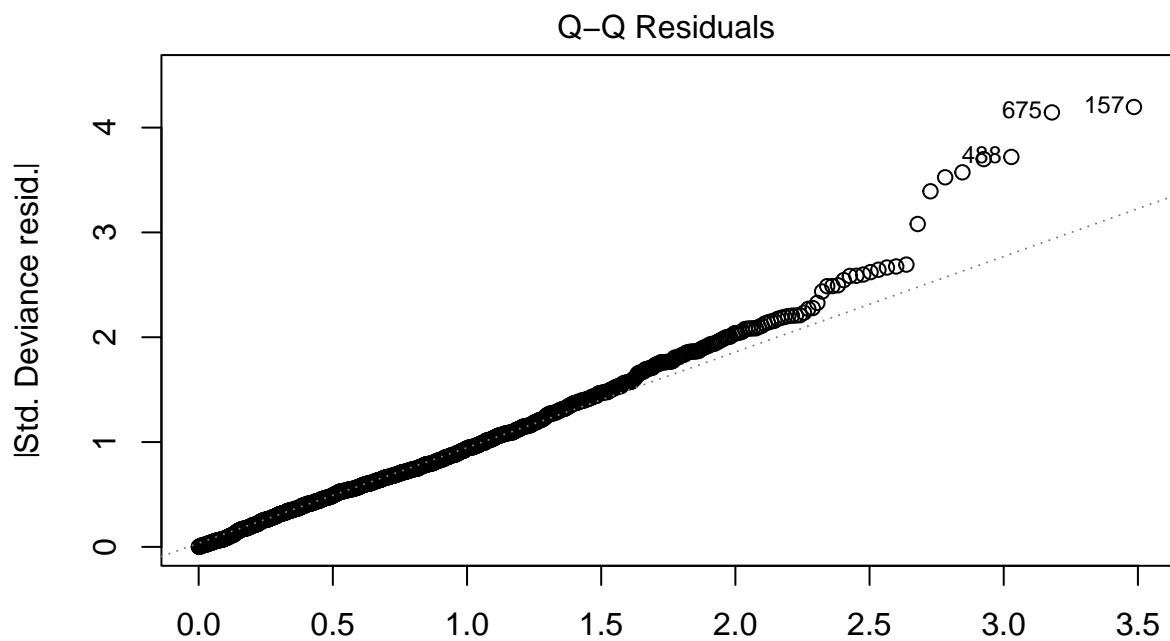
THE AIC we got from this lamda min approach was 353.7 which was a small jump from the base model . The VIF was acceptable with no variables exceeding the threshold of 10.

3.6 Model Evaluation

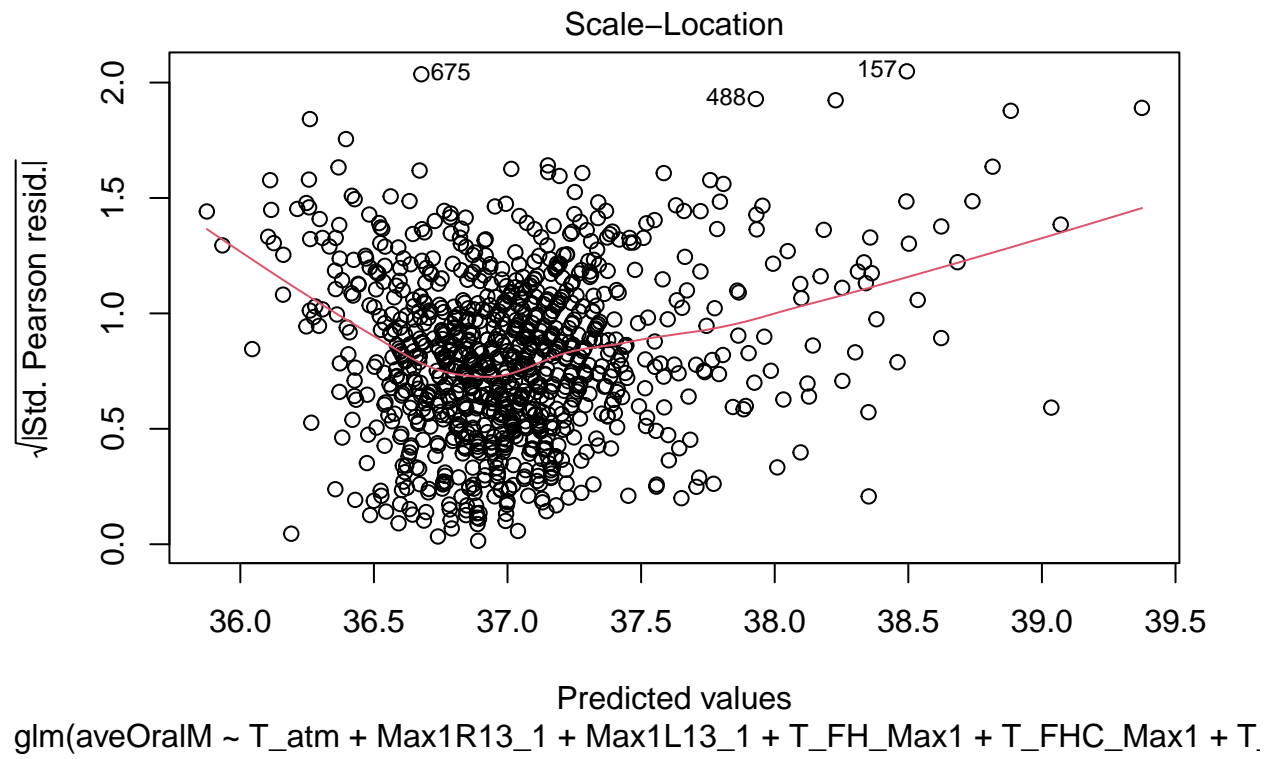
Linear model

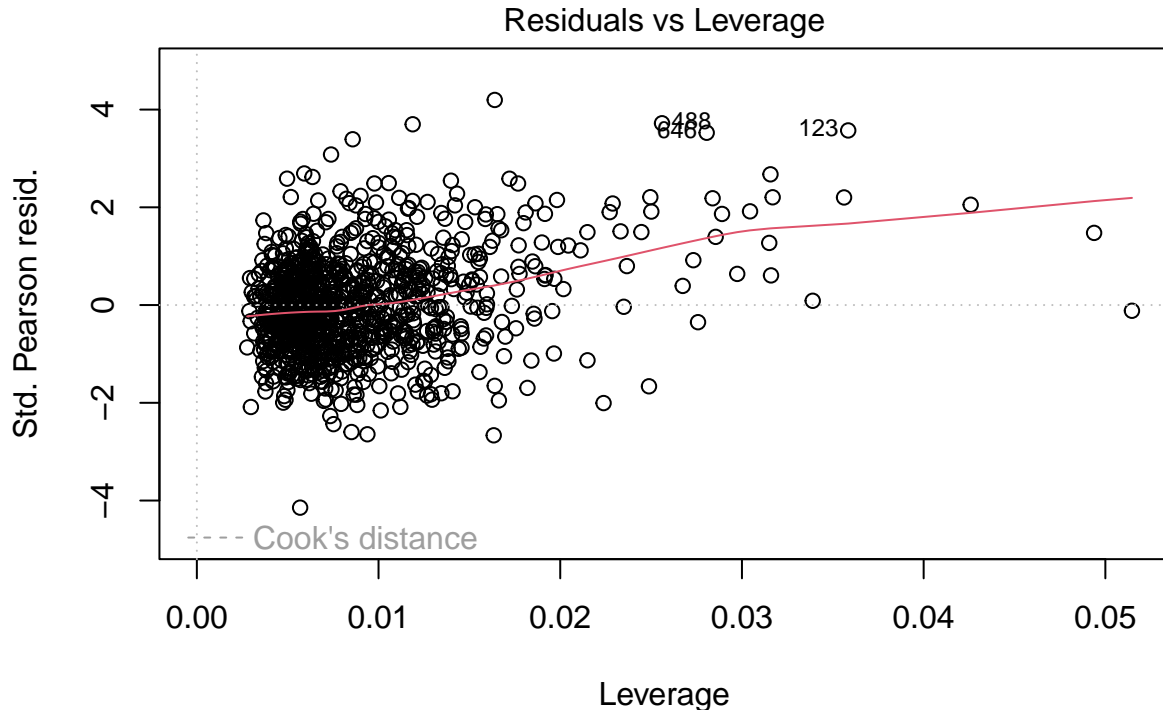


$\text{glm}(\text{aveOralM} \sim \text{T_atm} + \text{Max1R13_1} + \text{Max1L13_1} + \text{T_FH_Max1} + \text{T_FHC_Max1} + \text{T_}$



$\text{glm}(\text{aveOralM} \sim \text{T_atm} + \text{Max1R13_1} + \text{Max1L13_1} + \text{T_FH_Max1} + \text{T_FHC_Max1} + \text{T_}$





$\text{glm}(\text{aveOralM} \sim \text{T_atm} + \text{Max1R13_1} + \text{Max1L13_1} + \text{T_FH_Max1} + \text{T_FHC_Max1} + \text{T_}$

When we ran the model diagnostics, we found a non-linear, quadratic pattern in the Residuals vs Fitted plot, indicating that the model is not a great fit for this dataset.

The QQ residuals plot shows that the data is still right skewed with a heavy tail, which indicates non-normality and makes the Gaussian Linear model a bad fit for this dataset.

For the scale-location plot, we see that the model is not a great fit due to the quadratic nature of the best fit line. The quadratic nature of the best fit line shows that the model does not show a good spread, and violates the homoscedasticity assumption (violates constant variance).

In the Residuals vs Leverage plot, we did not see any high influence points, but there were three high leverage points at points 488, 646, 123.

Another observation found is that the data points were clustered around the 36.5 to 37.5 Degrees Celsius range.

Out of all the methods for variable reduction for LM, the backward elimination and forward elimination had the best AIC score of 272.6. The model had reduced the variables from 15 to 8. However, after performing the model diagnostic on the model, we found that it violates most of the assumptions of a linear model. Therefore, we conclude that the best reduced feature set cannot be explained well with a linear model.

Gamma Approach

With the gamma approach, we found that the best model was between the reduced feature set from ridge and backward/forward selection. The ridge model has an AIC around 350, and the other had approximately 250 in AIC. The VIFs were well within our desired range, which was below 10 for both models. The choice here would depend on our preferences. For instance, if we prefer a simpler model, opting for the ridge method would be a good option since it only includes 5 variables, while the other two selection methods resulted

in 8. However, due to the difference in AIC, it is also reasonable to choose the 8-variable model from the forward/backward selection method.

4. Discussion and Conclusion

For the Gamma model with the log link function, there is a huge difference between the null deviance and degrees of freedom, which potentially suggests a poor fit of the null model to the data. This potentially indicates that the model fails to capture the inherent variability in the response variable. When assessing how well the Gamma model performs, it's important to note the big difference between the null deviance and the residual deviance. For instance, when using ridge for our Gamma model, the null deviance was already very low at 0.189380 with 1016 degrees of freedom, whereas the residual deviance is significantly lower at 0.054498 with 1008 degrees of freedom. This large gap indicates that the null model, which includes no predictors, fits the data poorly. This suggests that we need to either rethink the current model's specification or consider different modeling strategies to improve fit.

Despite experimenting with exponential and logarithmic transformations of the variables, these methods did not result in any significant improvements, especially when compared to the performance of the gamma and linear models. However, the linear model didn't satisfy essential assumptions necessary for valid results, indicating problems with its use.

In conclusion, after testing two different modeling approaches and even attempting to simplify the models by reducing the number of features, none of the approaches produced satisfactory results. Maybe given more research, there could potentially be a model that would be a good fit for this data set.

Bibliography

Loo, C. (2020, March 29). Thermal scanners now in place at KLIA departure and arrival gates. *The Sun*.
https://thesun.my/local_news/thermal-scanners-now-in-place-at-klia-departure-and-arrival-gates-YJ2193203

Angeliki Kylili, Paris A. Fokaides, Petros Christou, Soteris A. Kalogirou (2014). Infrared thermography (IRT) applications for building diagnostics: A review. *Applied Energy*, 134, 531-549. ISSN 0306-2619.
<https://doi.org/10.1016/j.apenergy.2014.08.005>

data410 supu said something???