AN APPLICATION OF LINEAR AND LOGISTIC REGRESSION

on Children's Height Data in Sub-Saharan African Country

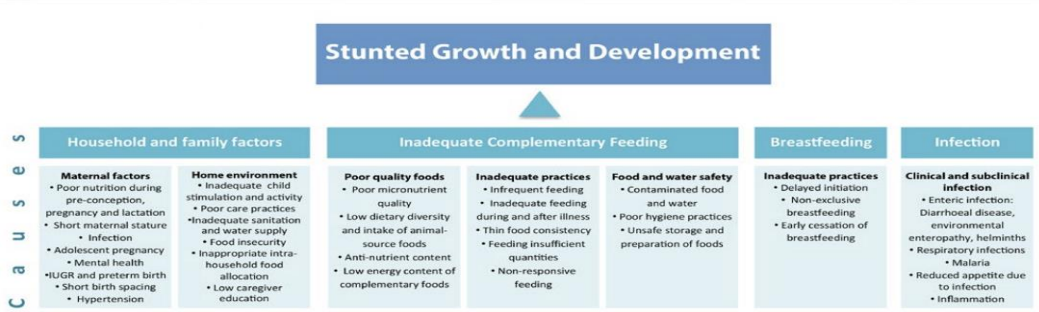
ALFIAN AFAN GHAFAR

2023

II. INTRODUCTION

A child's height is a valuable indicator to measure a child's health and nutrition well-being. Inadequate height might represent a deficiency of nutrition at an early age, the critical phase when the human body grows fastest. Furthermore, not only increases mortality and morbidity at an early age, but as explained by WHO (2016), this condition can bring negative outcomes later in adult life, including poor cognition and educational achievement, reduced work capacity and productivity, unachieved potential, and heightened risk of nutrition-related illnesses.

Children are classified as stunted when their length or height falls more than 2 standard deviations below the WHO child growth standards' median for the same age and gender (WHO, 2008). However, Onis & Branca (2016) remind us that it's crucial to recognize that, in reality, no strict boundaries separate one group of children as stunted and another as growing adequately. Instead, there exists a continuum of growth patterns. This means that the risk of stunting and the associated negative effects may not drastically change simply by crossing a specific threshold. Significant deterioration can also occur within the so-called 'normal' range.



(WHO, 2014)

Stunted child development can be attributed to various factors, including elements related to the household and family, as well as issues associated with complementary feeding, breastfeeding practices, and infections (WHO, 2014). Our study investigated the association between children's height (in z normal form) with the mother's physical condition at the time of giving birth, represented by weight and age, and also the nutrition from breastfeeding accepted by the infant. We also considered the parents' social circumstances, which are measured by wealth index, educational level, type of residence, and the number of living children in the family.

The data based on sub-Saharan African health survey data that hunger and malnutrition are some of the leading problems in child health in the region. It contains information on 3,799 children who were less than 5 years old at the time of the survey. The exploration of the data was divided into two sections based on methodology, the first was using linear regression with categorical variables followed by logistic regression by clustering the response variable into two groups, stunted and not.

III. OUTLINE OF METHODS

A. Linear Regression

Regression analysis was introduced by Galton (1822 - 1911) who studied the effect of a father's height on his son's height. Nowadays, regression terminology is applied to various forecasting types.

A simple linear regression model with one predictor (x) and one response (y) mathematically can be written: $y = \alpha + \beta x + \varepsilon$

Or with n predictor variables written: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon_i$

α: intercept; x: predictor; β_i: slope; ε_i: residual with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma_\varepsilon^2$

To obtain the best-fitting regression equation, parameter estimation aims to minimise the square of errors. One of the techniques for determining these parameters is known as Ordinary Least Squares (Walpole, 1982).

Assumption

There are several assumptions have to be fulfilled in using linear regressions, including:

a. There is a linear relationship between *y* and *x* in the population.
b. In the sample, for a fixed x_i, the observed y_i follows a normal distribution with a mean equal to the regression line and standard deviation lambda.
c. Lambda is constant for all *x*
d. The y_i are independent.

R squared

The coefficient of determination, denoted as R^2, represents the portion of variance accounted for by the model. It evaluates the goodness of fit, indicating a model's effectiveness in explaining outcomes within the context of linear regression. The calculation for R^2 is as follows: $R^2 = 1 - \frac{SSE}{SST}$ with $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ (Walpole,1982)

B. Logistic Regression

Sometimes linear regression cannot be applied when the dependent variable is dichotomous. In such cases, logistic regression is commonly used. Logistic regression and linear regression exhibit dissimilar model structures and underlying assumptions. Nevertheless, when acknowledging this distinction, the techniques applied in logistic regression follow the fundamental principles also seen in linear regression. Therefore, our approach to logistic regression draws inspiration from the same analytical techniques employed in linear regression (Homer, Lemeshov & Sturdivant, 2013).

Multiple logistic regression model is given by the equation:

$$g(x) = logit\,(\pi) = Log_e\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon_i$$

So, $$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}}$$

Wald Test

To identify the significance of variables, in logistic regression Wald test is used. Wald test is self is a comparison between the ratio of the maximum likelihood estimate of the slope parameter, β^1, to an estimate of its standard error.

$W = \frac{\widehat{\beta}_1}{\widehat{SE}(\widehat{\beta}_1)}$ (Homer, Lemeshov & Sturdivant, 2013).

Classification Table

If we use R-Square in linear regression to measure the predictive power of the model, in logistic regression, we can use a classification table to calculate the percentage of overall correct predictions (Kleinbaum & Klein, 2010).

Hosmer Lameshow

The goodness of fit in logistic regression can be provided by the Hosmer-Lameshow (HL) test. The method is based on logistic residual, called *m-asymptotics*. This term refers to clustering

observation data in covariate patterns and counting the successes and failures of the prediction. The Hosmer-Lameshow test is given by the equation:

$$H = \sum_{g=1}^{G} \left(\frac{(O_{1g} - E_{1g})^2}{E_{1g}} + \frac{(O_{0g} - E_{0g})^2}{E_{0g}} \right)$$ (Hilbe, 2009)

C. Model Selection

Not all initially included variables are significant to estimate the response variable. One step to gaining the best model in regression analysis is doing the model selection. It can be done by testing and comparing all potential models, unfortunately, this method is not always practical, considering some research contains many explanatory variables making it impossible to calculate all models. Fahrmei et al (2013) explained to make this phase easier, there are several methods could be used:

- Forward Selection: This selection begins with a basic model and then adds one most significant variable in each phase of iteration to the algorithm. The new variable is then contained in the model if the new model is significantly better than the previous one.
- Backward Elimination: The process begins with a complete model that includes all potential variables. At step gradually eliminates variables from the regression model if the new model after reduction is better or not significantly different in explaining the data.
- Stepwise Selection: This method is a combination of both forward selection and backward elimination. In each iteration, it allows for the possibility of either including or excluding a variable from the model.

IV. MODEL FITTING

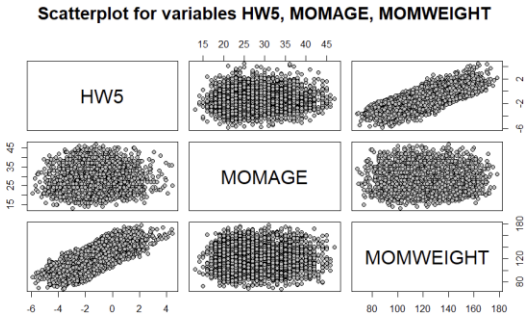
A. Variable Symbols

Symbol	Explanation		Symbol	Explanation
HW5	Height-for-Age z score		MOMAGE	Mother's age at birth of child
MOMWEIGHT	Mother's weight in pounds		SEX	Sex of child [male, female]
BREASTFEED	Duration of breastfeeding [still=still breastfeeding, <1=less than 1 year, b12=1-2 years, >2=more than 2 years]		MOMEDU	Mother's highest educational level [Ne=no education, Pri=primary, SecH=secondary or higher]
WEALTHIND	Household wealth index [pst=poorest, pr=poorer, m=middle, rc=rich, rcst=richest]		LIVCHN	Number of living children [<3=less than 3, B34=3-4, >5=5 or more]
RESIDENCE	Type of place of residence [urban, rural]			

B. Linear Regression

1. Linearity check

To begin with, it is essential to ensure that the relationship between continuous explanatory variables and the response follows a linear form. As we can observe from Picture 3.1, MOMWEIGHT



exhibits a positive linear relationship with HW5, while MOMAGE appears to be randomly distributed across x_i.

2. Hypothesis test

```
lm(formula = HW5 ~ MOMAGE + MOMWEIGHT + SEX + BREASTFEED_ + MOMEDU_ +
    WEALTHIND_ + RESIDENCE + LIVCHN_, data = dhs_data)
```

```
Residual standard error: 0.98 on 3783 degrees of freedom
Multiple R-squared:  0.5767,    Adjusted R-squared:  0.575
F-statistic: 343.5 on 15 and 3783 DF,  p-value: < 2.2e-16
```

$$H_0 = \beta_1 = \beta_2 = .. = \beta_p = 0 \text{=vs} \qquad H_1 = Not\ All\ \beta_i = 0$$

$$F = \frac{MSR}{MSE} = 513.3\ on\ 10\ and\ 3788\ DF, \quad p - value < 2.2e - 16$$

The decision is Reject H_0 means at least one of the explanatory variables is useful.

3. Model selection

Using the stepwise selection method in R with a p-value of 0.05, here is the output:

```
                        Stepwise Selection Summary
-------------------------------------------------------------------------------------
                               Added/                   Adj.
Step    Variable              Removed     R-Square     R-Square     C(p)        AIC         RMSE
-------------------------------------------------------------------------------------
   1    MOMWEIGHT             addition       0.570        0.570    49.3170   10678.5737    0.9861
   2    MOMEDU_               addition       0.572        0.572    27.9280   10659.3902    0.9833
   3    WEALTHIND_            addition       0.574        0.573    15.0870   10652.6052    0.9819
   4    BREASTFEED_           addition       0.575        0.574     5.5690   10647.0915    0.9808
-------------------------------------------------------------------------------------
> forward$model

Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Coefficients:
     (Intercept)         MOMWEIGHT         MOMEDU_NE        MOMEDU_Pri
         -7.79199           0.05601          -0.07452          -0.15560
   WEALTHIND_pst      WEALTHIND_pr       WEALTHIND_m      WEALTHIND_rc
         -0.18141          -0.16849          -0.10555          -0.04851
BREASTFEED_still     BREASTFEED_<1    BREASTFEED_B12
          0.05761          -0.02146          -0.06255
```

After conducting a stepwise selection process, the HOMEWEIGHT, MOMEDU, WEALTHIND, and BREASTFEED variables were added to the model as significant variables. Therefore, the new model includes these variables written by:

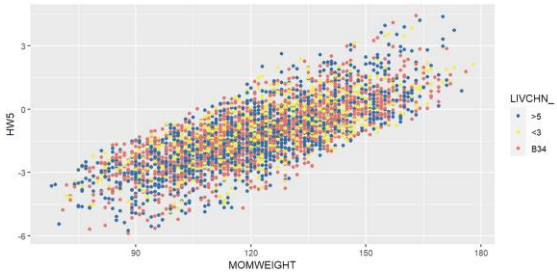
HW5 = -7.79 + 0.06MOMWEIGHT - 0.07MOMEDU_NE – 0.15MOMEDU_Pri – 0.18WEALTHININD_pst – 0.17WEALTHIDN_pr – 0.11WEALTHIDN_m – 0.049WEALTHIDN_rc + 0.06BREASTFEED_still – 0.02BREASTFEED_<1 – 0.06BREASTFEED_B12

4. Coefficient of determination

```
Call:
lm(formula = HW5 ~ MOMWEIGHT + BREASTFEED_ + MOMEDU_ + WEALTHIND_,
    data = dhs_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3090 -0.6353 -0.0096  0.6444  3.5252

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.7919919  0.1265590 -61.568  < 2e-16 ***
MOMWEIGHT       0.0560111  0.0008069  69.415  < 2e-16 ***
BREASTFEED_still  0.0576093  0.0645340   0.893  0.37208
BREASTFEED_<1    -0.0214612  0.0753574  -0.285  0.77582
BREASTFEED_B12   -0.0625463  0.0637437  -0.981  0.32655
MOMEDU_NE        -0.0745190  0.0566417  -1.316  0.18838
MOMEDU_Pri       -0.1556037  0.0447312  -3.479  0.00051 ***
WEALTHIND_pst    -0.1814103  0.0551889  -3.287  0.00102 **
WEALTHIND_pr     -0.1684858  0.0561548  -3.000  0.00271 **
WEALTHIND_m      -0.1055527  0.0573932  -1.839  0.06598 .
WEALTHIND_rc     -0.0485052  0.0576152  -0.842  0.39991
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9808 on 3788 degrees of freedom
Multiple R-squared:  0.5754,    Adjusted R-squared:  0.5742
F-statistic: 513.3 on 10 and 3788 DF,  p-value: < 2.2e-16
```

From the output $R^2 = 1 - \frac{SSE}{SST} = 0.5754$. It reveals that 57.54% of the variability of child height is explained by the regression model.

5. Interaction between Mother's Weight and Living Children



The scatterplot's Living Children to HW5 dots are randomly spread, making it difficult to identify interactions at a glance.

```
lm(formula = HW5 ~ MOMWEIGHT + BREASTFEED_ + MOMEDU_ + WEALTHIND
    MOMWEIGHT * LIVCHN_, data = dhs_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3433 -0.6318 -0.0069  0.6472  3.5293

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -8.308845   0.175798 -47.263  < 2e-16 ***
MOMWEIGHT                0.059986   0.001281  46.831  < 2e-16 ***
BREASTFEED_still         0.059283   0.064646   0.917 0.359179
BREASTFEED_<1           -0.023198   0.075218  -0.308 0.757786
BREASTFEED_B12          -0.061450   0.063644  -0.966 0.334348
MOMEDU_NE               -0.062783   0.057080  -1.100 0.271437
MOMEDU_Pri              -0.153026   0.044850  -3.412 0.000652 ***
WEALTHIND_pst           -0.174673   0.056230  -3.106 0.001908 **
WEALTHIND_pr            -0.162978   0.056955  -2.862 0.004239 **
WEALTHIND_m             -0.102430   0.057880  -1.770 0.076858 .
WEALTHIND_rc            -0.048407   0.057820  -0.837 0.402537
LIVCHN_<3                1.037877   0.270346   3.839 0.000126 ***
LIVCHN_B34               0.742614   0.219239   3.387 0.000713 ***
MOMWEIGHT:LIVCHN_<3     -0.008135   0.002189  -3.716 0.000205 ***
MOMWEIGHT:LIVCHN_B34    -0.005769   0.001787  -3.228 0.001259 **
---
```

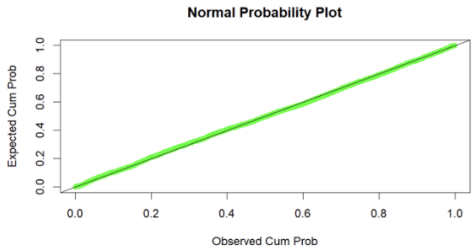
```
Analysis of Variance Table

Model 1: HW5 ~ MOMWEIGHT + BREASTFEED_ + MOMEDU_ + WEALTHIND_
Model 2: HW5 ~ MOMWEIGHT + BREASTFEED_ + MOMEDU_ + WEALTHIND_ + MOMWEIGHT *
    LIVCHN_
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3788 3644.2
2   3784 3625.9  4    18.367 4.7921 0.0007427 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

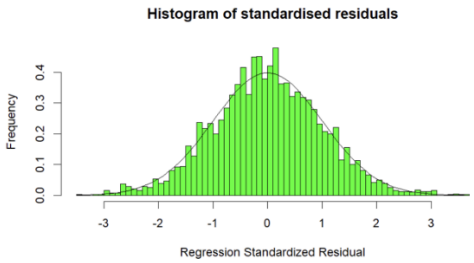
After adding the interaction between WEALTHIND and LIVCHN to the model, there was a significant variance difference (*p-value* < *0.05*), so both variables were included in the model.

HW5 = -8.31 + 0.06MOMWEIGHT - 0.06MOMEDU_NE – 0.15MOMEDU_Pri – 0.17WEALTHININD_pst – 0.16WEALTHIDN_pr – 0.10WEALTHIDN_m – 0.05WEALTHIDN_rc + 0.06BREASTFEED_still – 0.02BREASTFEED_<1 – 0.06BREASTFEED_B12 + 1.04 LIVCHN_<3 + 0.74 LIVCHN_B34 – 0.008MOMWEIGHT*LIVCHN_<3 – 0.006 MOMWEIGHT*LIVCHN_B34

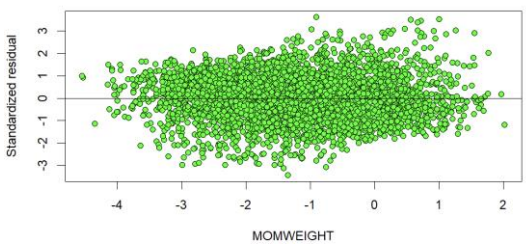
The new R^2 is equal to 0.5775, which means the new model with interaction able to represent the data slightly better.

6. Normal Assumption and Constant Variance





Standardized residual points lie along the diagonal line on a normal probability plot and the histogram of the residual graph shows a normal shape. These graphs confirm the normality assumption.



The residual plot appears satisfactory because the variance is relatively consistent throughout, and there are no concerning patterns. It proves that the assumption of constant variance has not been violated.

C. Logistic Regression

1. Convert height to binary model with height <-2 SD categorise as stunted.

```
Not-Stunting     Stunting
        2666         1133
```

there are 1.133 children in stunted group.

2. Model selection

Using stepwise selection with *p-value* = 0.05, here is the output of R

```
> qchisq(0.05,1,lower.tail=FALSE)
[1] 3.841459
> step.model <- model_L %>% stepAIC(trace = FALSE, k=3.84)
> summary(step.model)

Call:
glm(formula = HW5_1 ~ MOMWEIGHT + LIVCHN_ + MOMAGE + MOMEDU_,
    family = binomial, data = dhs_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 12.193720   0.560397  21.759  < 2e-16 ***
MOMWEIGHT   -0.105547   0.003662 -28.823  < 2e-16 ***
LIVCHN_<3   -0.730335   0.158990  -4.594 4.36e-06 ***
LIVCHN_B34  -0.530554   0.122415  -4.334 1.46e-05 ***
MOMAGE      -0.030630   0.009404  -3.257  0.00113 **
MOMEDU_NE    0.433402   0.160495   2.700  0.00693 **
MOMEDU_Pri   0.402225   0.137720   2.921  0.00349 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The new model after iteration written by the equation

$$g(x) = logit\,(\hat{\pi}) = 12.19 - 0.10MOMWEIGHT - 0.73LIVCHN\_<3 - 0.53LIVCHN\_B34 - 0.03MOMAGE + 0.43MOMEDU\_NE + 0.40MOMEDU\_Pri$$

3. Classification table

	dhs_data.HW5_L	glm_probs	glm_pred	dhs_data.MOMWEIGHT Mother's weight in pounds
10	Not-Stunting	0.4998315	Lower	111
3673	Stunting	0.4998315	Lower	111
3107	Not-Stunting	0.4999182	Lower	109
180	Stunting	0.5000102	Higher	108
888	Not-Stunting	0.5005400	Higher	106
2666	Stunting	0.5005400	Higher	106
1508	Not-Stunting	0.5013829	Higher	110
3016	Not-Stunting	0.5019127	Higher	108
2405	Stunting	0.5022026	Higher	107

Classification table devided predictions into two groups, predicted "stunted" if the probability is higher

```
glm_pred Not-Stunting Stunting
  Higher          263      699
  Lower          2403      434
```

There were 697 false predictions: 434 predicted stunted while not stunted and 263 predicted not stunted while stunted. The remaining 3,102 were predicted correctly, resulting in an overall accuracy rate of 81.65%.

4. Hosmer – Lameshov test

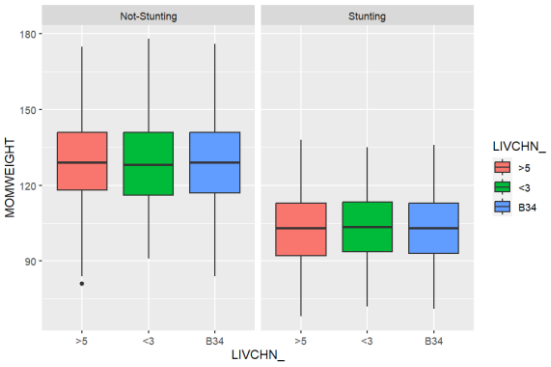
H_0: No difference between observed and fitted values (Model fits adequately)

H_1: There is difference (Model does not fit adequately)

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  dhs_data$HW5_1, fitted(logistic_model)
X-squared = 36.538, df = 28, p-value = 0.1294
```

With sub group 30, *p-value* is equal to 0.025, which mean model fits adequately.

5. Interaction between Mother's Weight and Living Children

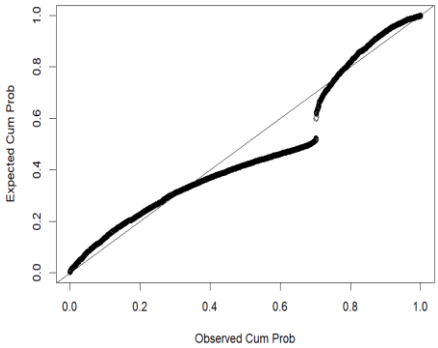


By glance at the boxplot graph, there is no difference of MOMWEIGHT in each group of LIVCHN

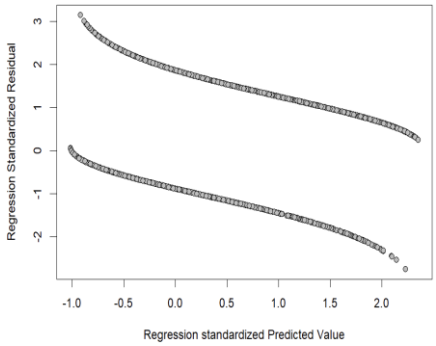
```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       12.553598   0.780792  16.078  < 2e-16 ***
MOMWEIGHT         -0.108677   0.005966 -18.215  < 2e-16 ***
LIVCHN_<3         -1.134233   1.156157  -0.981  0.32657
LIVCHN_B34        -1.185971   0.928662  -1.277  0.20158
MOMAGE            -0.030758   0.009424  -3.264  0.00110 **
MOMEDU_NE          0.430478   0.160528   2.682  0.00733 **
MOMEDU_Pri         0.401125   0.137546   2.916  0.00354 **
MOMWEIGHT:LIVCHN_<3  0.003562   0.010199   0.349  0.72693
MOMWEIGHT:LIVCHN_B34 0.005818   0.008172   0.712  0.47651
```

With the Chi-Squared test, we know that there is no evidence of an interaction between MOMWEIGHT and LIVCHN (no influence of MOMWEIGHT value to the effect of LIVCHN), so the interaction is not included in the model (*Chi-squared observation* = 0.48, *p-value* = 0.77 > 0.05). If we include the interaction in the model, the negative coefficient in the both level of LIVCHN and the positive coefficient in the both interaction suggest that the effect of LIVCHN on the probability of a stunted child is reduced for heavier mothers.

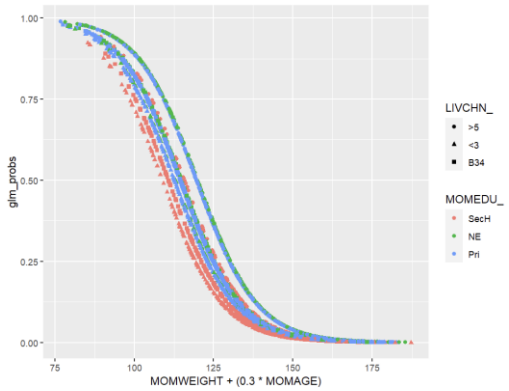
6. PP – Plot, Residual scatterplot, and S-Shaped probability plot



The PP-Plot graph compares the expected and the observed cumulative probability produced by the model. The disconnected lines are shaped because of significant changes of observed probability at some point.



There are two lines in the scatterplot of comparison between the standardized predicted value and to standardized residual representing a binary type of the model.



The probability of a child being stunted is higher when the mother is lighter and younger, as indicated by the negative effect of MOMWEIGHT and MOMAGE on the probability of HW5 < -2, which reverses the S-shaped curve. Additionally, a value of 0.3 is added to MOMAGE to reflect its relative weight as to MOMWEIGHT.

7. Marginal effect of MOMWEIGHT is $ME = \frac{\sum \pi(1-\pi)b_x}{n} = -0.0129$.

```
> mean(dhs_data$d_HW5_L <- glm_probs*(1-glm_probs)*(-0.105547))
[1] -0.01299718
```

V. RESULT AND COMPARISON

A. Linear Regression

Final linear regression model:

HW5 = -8.31 + 0.06MOMWEIGHT - 0.06MOMEDU_NE – 0.15MOMEDU_Pri – 0.17WEALTHININD_pst – 0.16WEALTHIDN_pr – 0.10WEALTHIDN_m – 0.05WEALTHIDN_rc + 0.06BREASTFEED_still – 0.02BREASTFEED_<1 – 0.06BREASTFEED_B12 + 1.04 LIVCHN_<3 + 0.74 LIVCHN_B34 – 0.008MOMWEIGHT*LIVCHN_<3 – 0.006 MOMWEIGHT*LIVCHN_B34

Considering it is a long regression model with several categorical variables, many of informations can be taken from it, for instance:

- Each additional pound of the mother's weight increases the expected height-for-age z score of the child by 0.06 points.
- A child born to a mother with no education is expected to have a height-for-age z score 0.06 points lower than a child born to a mother who completed secondary school or higher. Interestingly, the difference in expected height-z-score of a child from a mother with primary school education is relatively larger, at 0.15 points.
- The wealth index of the family plays a significant role in determining the height-for-age z score of a child. The children born from the "richest" category are expected to have the highest height-for-age z value. The expected deviation from the median value is 0.05 for the "richer", 0.10 for the "middle", 0.16 for the "poorer", and 0.17 for the poorest group.
- Mom with living children 4 or lower expected child height-for-age z score higher than mow with more than 4 children (1.04 if living children less than 3 and 0.74 if between 3-4). The gap expected lower with higher mom weight.
- Using all dummy level of categorical variables, a child from a mother with 160 pounds weight, mother secondary school last education, household wealth index as richest, breasfeeded more than 2 years, and has more than 3 siblings expected to has height-for-age z score of 1.29 (HW5 = -8.31+0.06(160) = 1.29).
- Here is a table of several example cases.

	Case 1	Case 2	Case 3	Case 4	Case 5
MOMWEIGHT	160	140	145	110	120
BREASTFEED	BREASTFEED_>2	BREASTFEED_B12	BREASTFEED_still	BREASTFEED_<1	BREASTFEED_>2
MOMEDU	MOMEDU_SecH	MOMEDU_SecH	MOMEDU_NE	MOMEDU_Pri	MOMEDU_NE
WEALTHIND	WEALTHIND_rcst	WEALTHIND_pst	WEALTHIND_pr	WEALTHIND_pst	WEALTHIND_pst
LIVCHN	LIVCHN_>5	LIVCHN_B34	LIVCHN_<3	LIVCHN_B34	LIVCHN_>5
Expected HW5	**1.29**	**0.72**	**1.27**	**-1.27**	**-1.34**
	Case 6	**Case 7**	**Case 8**	**Case 9**	**Case 10**
MOMWEIGHT	135	100	100	170	125
BREASTFEED	BREASTFEED_still	BREASTFEED_>2	BREASTFEED_<1	BREASTFEED_B12	BREASTFEED_>2
MOMEDU	MOMEDU_Pri	MOMEDU_Pri	MOMEDU_SecH	MOMEDU_NE	MOMEDU_SecH
WEALTHIND	WEALTHIND_m	WEALTHIND_rc	WEALTHIND_pr	WEALTHIND_rcst	WEALTHIND_rcst
LIVCHN	LIVCHN_B34	LIVCHN_B34	LIVCHN_<3	LIVCHN_>5	LIVCHN_>5
Expected HW5	**0.34**	**-1.77**	**-1.41**	**1.89**	**-2.91**

In case number 3, a child who was still breastfeeding from a mother weighing 145 pounds, and had less than 3 living children, came from a poor household and was expected to have a height-for-age z score index of 1.23.

- The R^2 of this model is 0.5775, which means 57.75% of the variance in the data is accounted for in this linear regression.

B. Logistic Regression

Final logistic regression model:

$g(x) = logit\ (\hat{\pi})$ = 12.19 - 0.10MOMWEIGHT – 0.73LIVCHN_<3 - 0.53LIVCHN_B34 – 0.03MOMAGE + 0.43MOMEDU_NE + 0.40MOMEDU_Pri

With model on odd scale: $\text{Odds} = \frac{\hat{\pi}}{1-\hat{\pi}} = e^{g(x)}$

And in probability scale: $\hat{\pi} = \frac{\exp\,(g(x))}{1+\exp\,(g(x))}$

This function gives us several information, including:

- For an extra pound of MOMWEIGHT, the logit probability of a stunted child decreases by 0.10, on average, decreases the odds of a stunted child by 9.51% (100*(exp(-0.10)-1)%).
- For every additional year of the mother's age, the average chance of her child being stunted decreases by a logit of 0.03, changing the odds of stunting multiplicatively by a factor of 0.96.
- The odds of stunted children are 1.54 times higher with uneducated mothers and 1.49 times higher with primary school mothers compared to those with mothers who completed secondary education. ($e^{0.43}$ = 1.54 & $e^{0.40}$ = 1.49).
- The more the siblings the child has, the more likely he/she to be stunted. The odds of a child being stunted are 2.075 times lower if he/she has less than three siblings and 1.699 times lower if he/she has 3 to 4 siblings compared to those who have more than 4 siblings ($e^{0.73}$ = 2.075 & $e^{0.53}$ =1.699). Furthermore, the odd to be stunted of a child with 3-4 siblings 1.50 greater than the odds a child with less than 3 siblings (($e^{12.19-0.59}$)/ ($e^{12.19-0.73}$) = 1.15)
- On average, the effect of one year of the mother's age is approximately a 1.30% decrease in the probability of having a stunted child.
- The estimated probability of a child from a 35-year-old mother who weighs 160 pounds, has more than four living children, and has completed university education being stunted is

  $$\hat{\pi} = \frac{\exp(12.19-(0.10*160)-(0.03*35))}{1+\exp(12.19-(0.10*160)-(0.03*35))} = 0.0077$$
- The estimated probability of a child from a 20-year-old mother who weighs 90 pounds, has 11 living children, and has no education being stunted is

  $$\hat{\pi} = \frac{\exp(12.19-(0.10*90)-(0.03*20)+0.43)}{1+\exp(12.19-(0.10*90)-(0.03*20)+0.43)} = 0.9535$$
- Here is the estimated probability table of some cases

	Case 1	Case 2	Case 3	Case 4	Case 5
MOMWEIGHT	160	140	145	110	120
MOMAGE	35	20	21	40	30
MOMEDU	MOMEDU_SecH	MOMEDU_NE	MOMEDU_SecH	MOMEDU_Pri	MOMEDU_NE
LIVCHN	LIVCHN_>5	LIVCHN_>5	LIVCHN_<3	LIVCHN_B34	LIVCHN_>5
Prob. Stunted	**0.007690876**	**0.953469525**	**0.024843544**	**0.465057055**	**0.430453776**
	Case 6	**Case 7**	**Case 8**	**Case 9**	**Case 10**
MOMWEIGHT	135	100	100	170	125
MOMAGE	18	24	45	27	33
MOMEDU	MOMEDU_Pri	MOMEDU_Pri	MOMEDU_SecH	MOMEDU_NE	MOMEDU_SecH
LIVCHN	LIVCHN_B34	LIVCHN_B34	LIVCHN_<3	LIVCHN_>5	LIVCHN_>5
Prob. Stunted	**0.820538481**	**0.005708921**	**0.291109827**	**0.005541132**	**0.900249511**

- From the model, there are 697 incorrect predictions: 434 for non-stunted as stunted, and 263 for stunted as non-stunted. Overall accuracy rate: 81.65%.

C. Comparison

- The main difference between linear regression and logistic regression is the form of the response variable. Linear regression's response variable is in numeric form, while logistic regression requires a binary format.
- Linear regression is used to identify the best-fitting line, while on the other hand logistic regression fits the line's values to a sigmoid curve ("S" shaped curve), allowing it to model non-linear relationships and classify data into discrete categories.

- Both models defined SEX and RESIDENCE as not significant variables, but there are some differences in the other variables (shown by the table). Also, linear regression counted the interaction between LIVCHN and MOMWEIGHT in the model, while logistic regression did not.

x	Liniar	Logistic
MOMAGE	Not-Significant	Significant
MOMWEIGHT	Significant	Significant
SEX	Not-Significant	Not-Significant
BREASTFEED	Significant	Not-Significant
MOMEDU	Significant	Significant
WEATHIND	Significant	Not-Significant
RESIDENCE	Not-Significant	Not-Significant
LIVCHN	Significant	Significant

- Linear regression provides greater detail in the outcome compared to logistic regression, which can only estimate the probability of a category. In the given case, logistic regression can estimate the probability of a child not being stunted (height-z-score>-2), but cannot predict the child's height-z-score.
- By using the linear model, we can get the prediction value of HW5 (called HW5_hat). Then by categorising HW5_hat <-2 as predicted stunted, classification table can be calculated.

	HW5_hat	
HW5	< -2	>= -2
< -2	706	427
>= -2	290	2376

Overall, there are 81.13% correct predictions from the linear model, which is slightly lower than the logistic model which can produce 81,65 correct predictions. It means in this case logistic regression has a better model in predicting stunted and non-stunted group.

VI. REFERENCES

De Onis, M., & Branca, F. (2016). Childhood stunting: a global perspective. *Maternal & child nutrition*, 12, 12-26.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B. (2013). Regression: Models, Methods and Applications. Germany: Springer Berlin Heidelberg.

Hilbe, J. (2009). Logistic regression models (Ser. Texts in statistical science). Chapman & Hall/CRC.

Hosmer, D. W., Lemeshow, S., Sturdivant, R. X. (2013). Applied Logistic Regression. United Kingdom: Wiley.

Kleinbaum, D. G., & Klein, M. (2010). Logistic regression: a self-learning text (3rd ed., Ser. Statistics for biology and health). Springer.

Walpole, R. E. (1982). Introduction to Statistics. Taiwan: Macmillan.

World Health Organization (2008). WHO Child Growth Standards: Training Course on Child Growth Assessment. Retrived from https://www.who.int/tools/child-growth-standards.

World Health Organization (2014). "Childhood stunting: challenges and opportunities. Retrived October 14, 2023, from https://iris.who.int/handle/10665/107026?&locale-attribute=fr.

World Health Organization (2016). “Childhood Stunting: Context, Causes and Consequences”. Retrived from https://www.who.int/publications/m/item/childhood-stunting-context-causes-and-consequences-framework.

VII. APPENDIX

1. Code

```
#Call the data
library(haven)
dhs_data <- read_sav("D:/STAT 6095 Regression Modelling/Coursework/demographic
and health survey.sav")
View(dhs_data)
colnames(dhs_data)
# Scatterplot HW5, MOMAGE, MOMWEIGHT
plot(dhs_data[,c("HW5", "MOMAGE", "MOMWEIGHT")],
     main = "Scatterplot for variables HW5, MOMAGE, MOMWEIGHT",
     bg = "grey", pch = 21)
#Replace Residence code
dhs_data$RESIDENCE <- dhs_data$RESIDENCE-1
#code SEX and residence with categorical variable labels
dhs_data$SEX<-factor(dhs_data$SEX, c(0,1), labels=c("Male", "Female"))
dhs_data$RESIDENCE<-factor(dhs_data$RESIDENCE, c(0,1), labels=c("Urban","Rural"))
# code as factor and assign labels
#BREASTFEED
dhs_data$BREASTFEED_<-factor(dhs_data$BREASTFEED, c(1,2,3,4),
                             labels=c("still","<1","B12",">2"))
dhs_data$BREASTFEED_ <- relevel(dhs_data$BREASTFEED_, ref = ">2")
#MOMEDU
dhs_data$MOMEDU_<-factor(dhs_data$MOMEDU, c(0,1,2),
                         labels=c("NE","Pri","SecH"))
dhs_data$MOMEDU_ <- relevel(dhs_data$MOMEDU_, ref = "SecH")
#WEALTHIND
dhs_data$WEALTHIND_<-factor(dhs_data$WEALTHIND, c(1,2,3,4,5),
labels=c("pst","pr","m","rc","rcst"))
dhs_data$WEALTHIND_<- relevel(dhs_data$WEALTHIND_, ref = "rcst")
#LIVCHN
dhs_data$LIVCHN_<-factor(dhs_data$LIVCHN, c(1,2,3),
                         labels=c("<3","B34",">5"))
dhs_data$LIVCHN_<- relevel(dhs_data$LIVCHN_, ref = ">5")

#LINEAR REGRESSION
model <- lm(formula = HW5 ~ MOMAGE+ MOMWEIGHT+ SEX+ BREASTFEED_
            + MOMEDU_+ WEALTHIND_+ RESIDENCE+ LIVCHN_,
            data = dhs_data)
summary(model)
# stepwise forward regression
library(olsrr)
```

```
stepwise<-ols_step_both_p(model,prem = 0.05,pent=0.05)
stepwise
stepwise$model
model_new <- lm(formula = HW5 ~ MOMWEIGHT+ BREASTFEED_
              + MOMEDU_+ WEALTHIND_,
              data = dhs_data)
summary(model_new)
#interaction between MOMWEIGHT and LIVCHN
#Scatterplot
library(ggplot2)
cols <- c("#1170AA", "#F2EF10", "#EF6F6A")
ggplot(dhs_data, aes(x = MOMWEIGHT, y = HW5, color = LIVCHN_)) +
  geom_point() +
  scale_color_manual(values = cols)
#Make a model with interaction
model_Intr <- lm(HW5~MOMWEIGHT+ BREASTFEED_
               + MOMEDU_+ WEALTHIND_+ LIVCHN_
               + MOMWEIGHT*LIVCHN_, data = dhs_data)
summary(model_Intr)
anova(model_new, model_Intr) #Significant
# produce residual vs. fitted plot
res <- resid(model_Intr)
plot(fitted(model_Intr), scale(res), xlab = "MOMWEIGHT", ylab =
       "Standardized residual", bg = "green", pch = 21)
abline(0,0)
# PP plot
# get probability distribution for residuals
probDist <- pnorm(scale(res))
# create PP plot
plot(ppoints(length(scale(res))), sort(probDist),col="green",
     xlab = "Observed Cum Prob",    ylab = "Expected Cum Prob",
     main="Normal Probability Plot")
abline(0,1)
# Histogram
# create residuals histogram
hist(scale(res), freq = FALSE, breaks=60, col="green",
     xlab= "Regression Standardized Residual",
     ylab = "Frequency",
     main="Histogram of standardised residuals")
curve(dnorm, add = TRUE)
#LOGISTIC REGRESSION
#Transform HW5 into binary
```

```
# As a 0/1 variable
dhs_data$HW5_1 <- as.numeric(dhs_data$HW5 < -2)
dhs_data$HW5_L<-factor(dhs_data$HW5_1, c(0,1), labels=c("Not-Stunting",
"Stunting"))
table(dhs_data$HW5_L, exclude = NULL)
# fit logistic regression
model_L <- glm(formula = HW5_1 ~ MOMWEIGHT+ LIVCHN_+ MOMAGE
              + MOMEDU_+ WEALTHIND_+ SEX+ RESIDENCE+ BREASTFEED_,
              data = dhs_data, family = binomial)
summary(model_L)
#Run Stepwise for logistic regression
nothing <- glm(HW5_L ~ 1,family=binomial,data = dhs_data)
step.model <- step(nothing,scope = list(upper=model_L),
                   direction="both",test="Chisq", trace = F,
                   k=qchisq(0.05,1,lower.tail=FALSE))
summary(step.model)
logistic_model <- step.model
#Interaction between MOMWEIGHT and LIVCHN
#Boxplot
cols <- c("#1170AA", "#F2EF10", "#EF6F6A")
ggplot(dhs_data, aes(x = HW5_L, y = MOMWEIGHT, fill = HW5_L)) +
  geom_boxplot() +
  scale_color_manual(values = cols) +
  facet_wrap(~LIVCHN_)
# fit logistic regression with interaction
model_Lint<-glm(formula=HW5_1~MOMWEIGHT+ LIVCHN_+ MOMAGE
            + MOMEDU_+ MOMWEIGHT * LIVCHN_,
            data = dhs_data, family = binomial)
summary (model_Lint)
# Likelihood ratio test / Anova
anova(model_Lint,test="Chisq") #Not Significant
summary (glm(HW5_1~MOMWEIGHT+ LIVCHN_+ MOMAGE
             + MOMEDU_+ MOMWEIGHT * LIVCHN_,
             data = dhs_data, family = binomial)) #See with interaction
# Classification Table
glm_probs <- predict(logistic_model,type = "response")
glm_pred <- ifelse(glm_probs > 0.5, "Higher", "Lower")
# print table
df_ct<-data.frame(dhs_data$HW5_L, glm_probs, glm_pred, dhs_data$MOMWEIGHT)
head(df_ct)
Csum<-table(glm_pred, dhs_data$HW5_L)
Csum
```

```
#Calculate accuracy rate
predicted <- ifelse(glm_probs > 0.5, "Stunting", "Not-Stunting")
mean(predicted == dhs_data$HW5_L)
# Hosmer-Lemeshow test
library(ResourceSelection)
HL <- hoslem.test(dhs_data$HW5_1, fitted(logistic_model), 30)
HL
cbind(HL$observed, HL$expected)
# compute the marginal effect of MOMWEIGHT
dhs_data$d_HW5_L <- glm_probs*(1-glm_probs)*(-0.105547)
mean(dhs_data$d_HW5_L)
mean(dhs_data$d_HW5_L <- glm_probs*(1-glm_probs)*(-0.105547))
```

2. Anova test to Logictic Model Interaction

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: HW5_1

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                3798     4629.9
MOMWEIGHT          1  1658.75      3797     2971.2 < 2.2e-16
LIVCHN_            2    22.01      3795     2949.2 1.665e-05
MOMAGE             1    16.88      3794     2932.3 3.985e-05
MOMEDU_            2     9.51      3792     2922.8  0.008598
MOMWEIGHT:LIVCHN_  2     0.51      3790     2922.3  0.774476

NULL
MOMWEIGHT          ***
LIVCHN_            ***
MOMAGE             ***
MOMEDU_            **
MOMWEIGHT:LIVCHN_
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```