Malco Oway Tolentino

Regression Analysis Project

Summer 2014

This project is divided into the following sections:

A. Main Objective
B. Specific Objectives
C. The Data
D. Assumptions and Limitations
E. Data Analysis
F. The Model
G. Data Transformation
H. Model Fitting
I. Model Diagnostic Checks
J. Model Interpretation and Forecasting

A. Main Objective

This project aims to model the 2010 Philippine population by age and gender groups using the techniques of regression analysis.

B. Specific Objectives

The specific objectives are:

- Describe the 2010 Philippine population data using graphs.
- Perform necessary data transformations suitable for the development of regression models and identify the key parameters.
- Perform diagnostic checks by testing the significance of the parameters and analysing the residuals.
- Use the derived model to forecast the population from a given age and gender.

C. The Data

The data used in this project is the 2010 Philippine population by age and gender from the Philippine Statistics Authority – National Statistical Coordination Board. The primary source is the census from National Statistics Office. The data is available for public use and can be accessed thruhttp://www.nscb.gov.ph/secstat/d_popn.asp.

Table 1. 2010 Philippine Population by Age Group and Gender (Source: National Statistics Office)

Age Group	Both Sexes	Male	Female
Philippines	**92,097,978**	**46,459,318**	**45,638,660**
Under 5	10,231,648	5,291,880	4,939,768
5-9	10,317,657	5,329,978	4,987,679
10-14	10,168,219	5,230,893	4,937,326
15-19	9,676,359	4,914,379	4,761,980
20-24	8,370,398	4,229,958	4,140,440
25-29	7,390,062	3,719,437	3,670,625
30-34	6,744,028	3,419,039	3,324,989
35-39	5,990,108	3,037,467	2,952,641
40-44	5,450,679	2,761,377	2,689,302
45-49	4,664,537	2,354,757	2,309,780
50-54	3,883,630	1,945,258	1,938,372
55-59	2,980,350	1,470,861	1,509,489
60-64	2,224,105	1,061,324	1,162,781
65-69	1,495,115	678,782	816,333
70-74	1,140,951	491,491	649,460
75-79	705,977	285,693	420,284
80-84	393,387	145,686	247,701
85 and over	270,768	91,058	179,710

D. Assumptions and Limitations

We will limit the modelling of the population with the following assumptions:

- The explanatory variables will be age and gender.

- The gender will implicitly be the conditional variable.

- Since the ages are tabulated based on the age group (4-year interval, except for the age group "85 and over"), for modelling purposes, the age variable shall take the midpoint of the interval. That is, if x is the age variable for the age group, then

$$x = \frac{Age_{Lower\ Bound} + Age_{Upper\ Bound}}{2}$$

where $Age_{Lower\ Bound}$ and $Age_{Upper\ Bound}$ are the lower and upper bound of the age group.

For example, for the age group "Under 5", the age variable would take the value $x = (0+4)/2 = 2$. Similarly, the value of the age variable for the age group "80-84" would be $x = (80+84)/2 = 82$. Note that the upper bound of the age group "85 and over" is not explicitly defined in the data. For this reason, the data will be truncated to exclude the last age group and hence, for modelling purposes, the age data shall only include the values from 0 to 84.

- Arbitrary limiting ages will be assumed for males and females (Refer to Data Analysis section.)

- The dependent variable will be the Philippine population for 2010.

E. Data Analysis

The conditional plots of the 2010 Philippine population are shown below (where gender is the conditioning variable).The points in the graphs are connected to emphasize the trend. Figure 3 shows the male and female populations on the same graph.

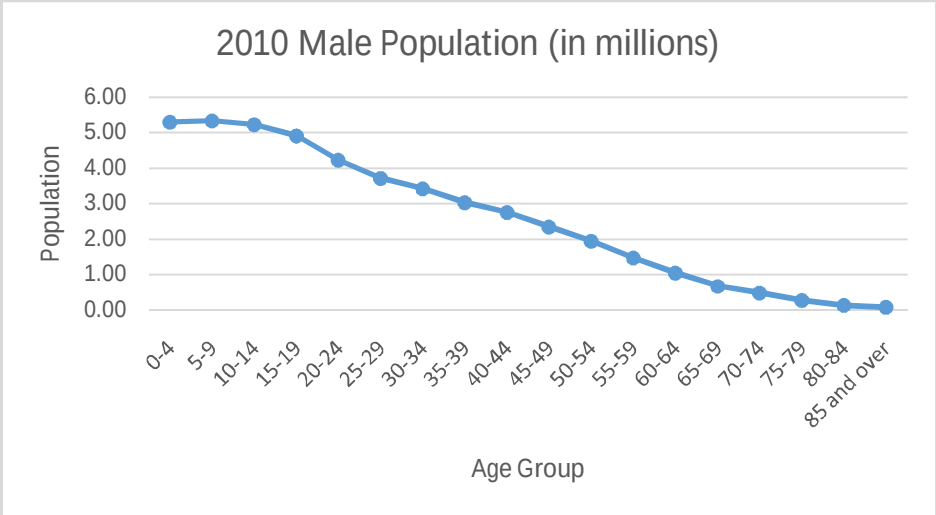

Figure 1. Graph of male population (in millions) by age group

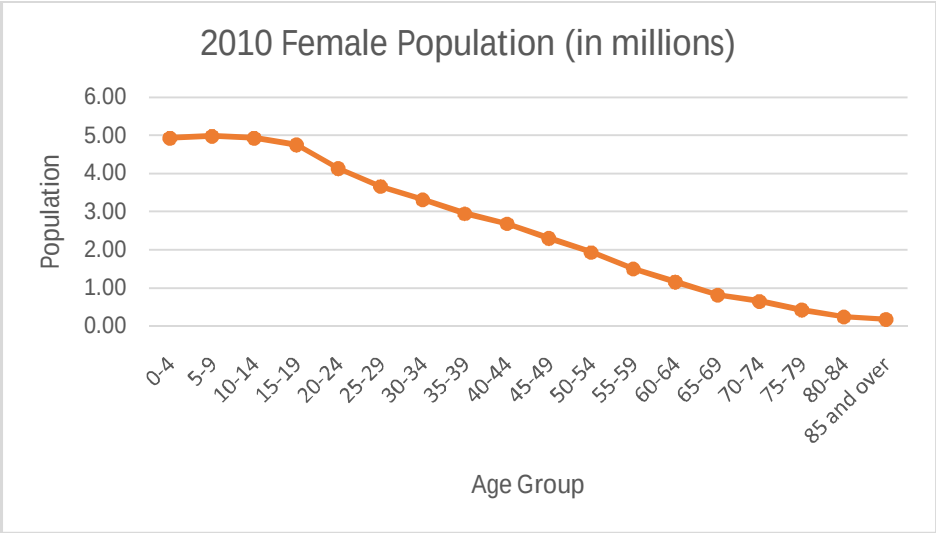

Figure 2. Graph of female population (in millions) by age group

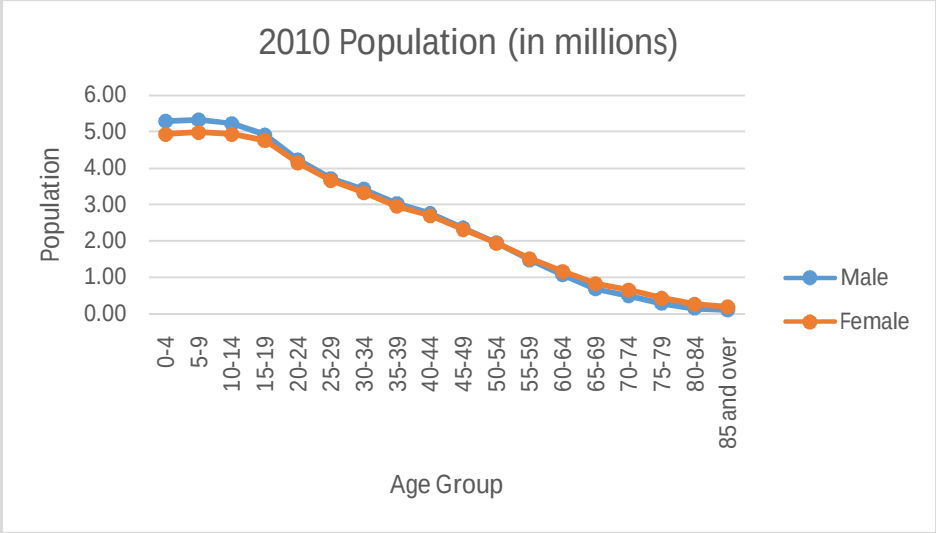

Figure 3. Graph of male and female population (in millions) by age group

The graphs show some inherent characteristics of population data, or more accurately human mortality rates. In any population, there is an "initial" value that approaches to zero as the age increases. (In fact, this is true even for the aggregate population.) The population also tends to decrease slower at younger ages (below 20) and older ages (starting at age 70) than at the middle ages (ages 20 to 69). Furthermore, the rate of population decrease is fairly constant at middle ages 20 to 69, because the population trend at those ages is approximately linear.

Across gender, there is an apparent gap between males and females at younger and older ages than at the middle ages. The male population is higher than female at ages roughly 0 to 19 while lower than female at ages around 60 and up, which depicts that "women outlive men". At the middle ages (20 to 59), there appears to be a 50% male and 50% female distribution in the 2010 Philippine population.

Since the population will be zero at some specified age, we will assume a limiting age for each gender. Note that the weighted average age (WAA) of males is 25.92 and the WAA for females is 26.96 (the weights applied being the population of each gender). We will arbitrarily assign a limiting age of 100 for males and since the WAA of females is 4% higher than males, we will arbitrarily assign a limiting age of 104 for females.

F. The Model

The Base Model

Let $P(x)$ be the population at age x. The proposed model for the 2010 Philippine population is given by

$$P(x) = C_M \exp\left(\frac{\beta_M x}{100 - x}\right) + C_F \exp\left(\frac{\beta_F x}{104 - x}\right) \qquad \text{(Eq-1)}$$

where,

$\exp(x)$ is the exponential function e^x,

C_M and C_F are constants corresponding to the initial population of males and females, respectively

β_M and β_M are constant 'scale' parameters for males and females, respectively.

The assumed limiting ages for males and females are 100 and 104, respectively. The "male" and "female" terms of Eq-1 are, respectively,

$$P_M(x) = C_M \exp\left(\frac{\beta_M x}{100 - x}\right) \text{ and} \qquad \text{(Eq-2)}$$

$$P_F(x) = C_F \exp\left(\frac{\beta_F x}{104 - x}\right), \qquad \text{(Eq-3)}$$

To determine the reasonability of this model, we will investigate its properties, specifically Eq-2 and Eq-3.

- For males (i.e., in Eq-2), as $x \to 0$, $P_M(x) \to C_M$. Similarly, for females (i.e., in Eq-3), as $x \to 0$, $P_F(x) \to C_F$. That is, there is an "initial" population of males and females.

- For males, as $x \to 100$, $P_M(x) \to 0$. Similarly, for females, as $x \to 104$, $P_F(x) \to 0$. This implies that the population vanishes as the individuals approach their limiting age.

- For males, the expression $\frac{x}{100 - x}$ in the $\exp$ function is always positive since the most value that x can attain is the limiting age of 100. The same is true for females, the expression $\frac{x}{104 - x}$ in the $\exp$ function is always positive. Hence, if $\beta_M < 0$ and $\beta_F < 0$, then Eq-2 and Eq-3 are both decreasing functions of x. We will show in Model Fitting section that the scale parameters of Eq-2 and Eq-3 are indeed negative.

The above properties of the model coincide with the same properties of the population data. The fitted model (including the parameter estimates) will be discussed in the Model Fitting section.

The "Linearized" Model

If an error term is included in Eq-2 and Eq-3, the model becomes

$$P_M(x) = C_M \exp\left(\frac{\beta_M x}{100 - x} + \varepsilon_M\right) \qquad \text{(Eq-4)}$$

$$P_F(x) = C_F \exp\left(\frac{\beta_F x}{100 - x} + \varepsilon_F\right) \quad \text{(Eq-5)}$$

$$P(x) = P_M(x) + P_F(x) \quad \text{(Eq-6)}$$

where $\varepsilon_M \sim Normal\left(0, \sigma^2_{\varepsilon_M}\right)$ and $\varepsilon_F \sim Normal\left(0, \sigma^2_{\varepsilon_F}\right)$, i.e., the error term is normally distributed with zero mean and constant variance.

To apply the linear regression analysis, the relationship between the dependent variable and the independent variables should necessarily be linear. Since Eq-4 and Eq-5 are non-linear functions, we need to transform the model to arrive at it's "linear" form. By taking the natural logarithm of both sides of Eq-4, we obtain

$$\ln\left(P_M(x)\right) = \ln\left(C_M\right) + \frac{\beta_M x}{100 - x} + \varepsilon_M \quad \text{(Eq-7)}$$

Applying the substitutions

$$Y_M = \ln\left(P_M(x)\right), \quad \text{(Eq-8)}$$

$$\alpha_M = \ln\left(C_M\right), \quad \text{(Eq-9)}$$

$$\text{and } X_M = \frac{x}{100 - x}, \quad \text{(Eq-10)}$$

Eq-7 will be transformed into the linear model

$$Y_M = \alpha_M + \beta_M X_M + \varepsilon_M . \quad \text{(Eq-11)}$$

Similarly, by taking the natural logarithm of both sides of Eq-5, and applying the substitutions

$$Y_F = \ln\left(P_F(x)\right), \quad \text{(Eq-12)}$$

$$\alpha_F = \ln\left(C_F\right), \quad \text{(Eq-13)}$$

$$\text{and } X_F = \frac{x}{104 - x}, \quad \text{(Eq-14)}$$

Eq-5 can be transformed into the linear model

$$Y_F = \alpha_F + \beta_F X_F + \varepsilon_F . \quad \text{(Eq-15)}$$

G. Data Transformation

To fit the data into the linear models Eq-11 and Eq-15, the population data will be transformed according to Eq-8 and Eq-12 for the males and females dependent variables, respectively. The equations Eq-10 and Eq-14 will be used to transform the age variable for the explanatory variables for males and females, respectively.

The original and transformed variables are summarized in Table 2 and Table 3. The corresponding graphs are shown in Figure 4 and Figure 5 with the apparent linear relationship between the dependent and independent variables.

Table 2. Original and Transformed 2010 Philippine Male Population Data

Age Group	Age Variable (x)	Gender	Population	X_M	Y_M
Under 5	2	Male	5,291,880	0.02	15.48
5-9	7	Male	5,329,978	0.08	15.49
10-14	12	Male	5,230,893	0.14	15.47
15-19	17	Male	4,914,379	0.20	15.41
20-24	22	Male	4,229,958	0.28	15.26
25-29	27	Male	3,719,437	0.37	15.13
30-34	32	Male	3,419,039	0.47	15.04
35-39	37	Male	3,037,467	0.59	14.93
40-44	42	Male	2,761,377	0.72	14.83
45-49	47	Male	2,354,757	0.89	14.67
50-54	52	Male	1,945,258	1.08	14.48
55-59	57	Male	1,470,861	1.33	14.20
60-64	62	Male	1,061,324	1.63	13.88
65-69	67	Male	678,782	2.03	13.43
70-74	72	Male	491,491	2.57	13.11
75-79	77	Male	285,693	3.35	12.56
80-84	82	Male	145,686	4.56	11.89

Table 3. Original and Transformed 2010 Philippine Female Population Data

Age Group	Age Variable (x)	Gender	Population	X_F	Y_F
Under 5	2	Female	4,939,768	0.02	15.41
5-9	7	Female	4,987,679	0.07	15.42
10-14	12	Female	4,937,326	0.13	15.41
15-19	17	Female	4,761,980	0.20	15.38
20-24	22	Female	4,140,440	0.27	15.24
25-29	27	Female	3,670,625	0.35	15.12
30-34	32	Female	3,324,989	0.44	15.02
35-39	37	Female	2,952,641	0.55	14.90
40-44	42	Female	2,689,302	0.68	14.80
45-49	47	Female	2,309,780	0.82	14.65
50-54	52	Female	1,938,372	1.00	14.48
55-59	57	Female	1,509,489	1.21	14.23
60-64	62	Female	1,162,781	1.48	13.97
65-69	67	Female	816,333	1.81	13.61
70-74	72	Female	649,460	2.25	13.38
75-79	77	Female	420,284	2.85	12.95
80-84	82	Female	247,701	3.73	12.42

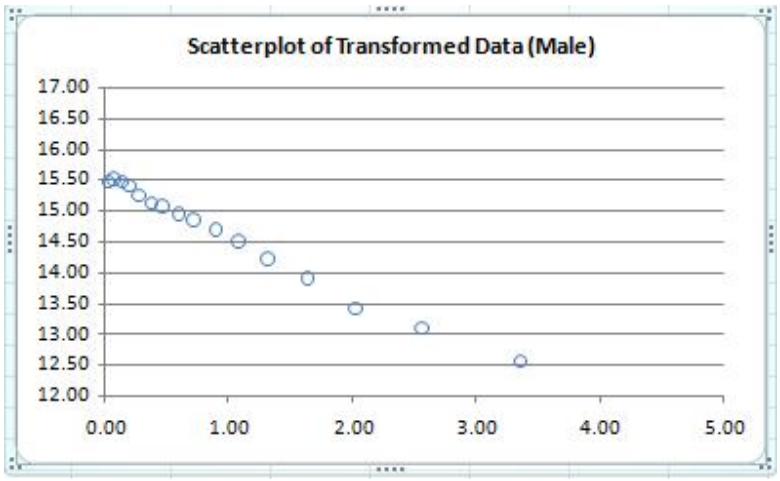

Figure 4. Scatterplot of the transformed male population (y-axis) and age (x-axis)

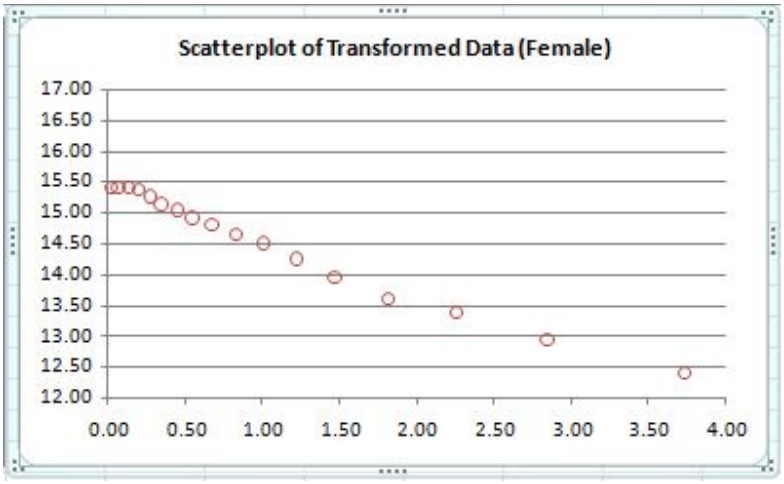

Figure 5. Scatterplot of the transformed female population (y-axis) and age (x-axis)

H. Model Fitting

(Refer to the Excel regression template for all the required computations)

The estimated parameters of the linear model Eq-11 are $\hat{\beta}_M = -0.8523$ and $\hat{\alpha}_M = 15.4445$, and the fitted linear regression model is

$$\hat{Y}_M = 15.4445 - 0.8523 X_M \qquad \text{(Eq-16)}$$

In terms of the male population model Eq-4,

$$\hat{P}_M(x) = 5{,}098{,}652 \exp\left(-\frac{0.8523x}{100-x}\right) \qquad \text{(Eq-17)}$$

The constant factor is determined from Eq-9: since $\hat{\alpha}_M = 15.4445 = \ln(C_M)$, then $C_M = e^{15.4445} = 5{,}098{,}652$. The tabular *t*-value with $n-2 = 15$ degrees of freedom at $\alpha = 1\%$ level of significance is 3.2860. The standard error of the estimate of the slope parameter of the male linear model is 0.0283. Then the 99% confidence interval for the slope parameter is determined as

$$CI = \hat{\beta}_M \pm 3.2860\,(0.0283) = -0.8523 + 3.2860\,(0.0283) = [-0.9454, -0.7592]$$

The standard error of the intercept parameter of the male linear model is 0.0488. Then the 99% confidence interval of the intercept parameter is determined as

$$CI = \hat{\alpha}_M \pm 3.2860\,(0.0488) = 15.4445 + 3.2860\,(0.0488) = [15.2841, 15.6048].$$

Similarly, the estimated parameters of the linear model Eq-15 are $\hat{\beta}_F = -0.8679$ and $\hat{\alpha}_F = 15.4052$, and the fitted linear regression model is

$$\hat{Y}_F = 15.4052 - 0.8679\,X_F \qquad \text{(Eq-18)}$$

In terms of the female population model Eq-5,

$$\hat{P}_F(x) = 4{,}902{,}279\,\exp\left(-\frac{0.8679x}{104-x}\right) \qquad \text{(Eq-19)}$$

The constant factor is determined from Eq-13: since $\hat{\alpha}_F = 15.4445 = \ln(C_F)$, then $C_F = e^{15.4052} = 4{,}902{,}279$. The standard error of the estimate of the slope parameter of the female linear model is 0.0277. Then the 99% confidence interval for the slope parameter is determined as

$$CI = \hat{\beta}_F \pm 3.2860\,(0.0277) = -0.8679 + 3.2860\,(0.0277) = [-0.9588, -0.7770]$$

The standard error of the intercept parameter of the female linear model is 0.0408. Then the 99% confidence interval of the intercept parameter is determined as

$$CI = \hat{\alpha}_F \pm 3.2860\,(0.0408) = 15.4052 + 3.2860\,(0.0408) = [15.2713, 15.5392].$$

The fitted model for the 2010 Philippine Population data is given by

$$\hat{P}(x) = \hat{P}_M(x) + \hat{P}_F(x)$$

$$\hat{P}(x) = 5{,}098{,}652\,\exp\left(-\frac{0.8523x}{100-x}\right) + 4{,}902{,}279\,\exp\left(-\frac{0.8679x}{104-x}\right) \qquad \text{(Eq-20)}$$

Linear Model for Log Male Population

The total sum of squares, regression sum of squares and the residual sum of squares for the fitted linear model for males are as follows:

$$TSS_M = 19.3470,\ Reg\,SS_M = 19.0316, \text{ and } RSS_M = 0.3154$$

The square of the correlation coefficient is $R_M^2 = Reg\,SS_M / TSS_M = 0.9837$, while the Adjusted $R_M^2 = 0.9826$. That is, 98.26% of the variability between the dependent variable Y_M and the explanatory variable X_M is explained by the fitted linear model for males.

The standard error of the regression and the estimated regression parameters are:

$$\hat{\sigma}_{\varepsilon_M} = \sqrt{RSS_M/(n-2)} = 0.1450,\ SE\left(\hat{\beta}_M\right) = 0.0283, \text{ and } SE\left(\hat{\alpha}_M\right) = 0.0488.$$

The analysis that follows tests the significance of the regression parameters at level of significance $\alpha = 1\%$ using the two-tailed *t*-test. (Note: In the context of hypothesis testing, α refers to the level of significance not the intercept parameter of the linear regression model.). The tabular *t*-value with $n-2=15$ degrees of freedom at $\alpha = 1\%$ level of significance is $t_{\alpha/2} = 3.2860$.

Null hypothesis: $H_0 : \beta_M = 0$

t-statistic: $$t_0 = \frac{\hat{\beta}_M - 0}{SE\left(\hat{\beta}_M\right)} = \frac{-0.8523}{0.0283} = -30.09$$

Since $|t_0| > t_{\alpha/2}$ at $\alpha = 1\%$, the null hypothesis $H_0 : \beta_M = 0$ is rejected. Hence, the slope parameter of the male linear model is significant.

Null hypothesis: $H_0 : \alpha_M = 0$

t-statistic: $$t_0 = \frac{\hat{\alpha}_M - 0}{SE\left(\hat{\alpha}_M\right)} = \frac{15.4445}{0.0488} = 316.48$$

Since $|t_0| > t_{\alpha/2}$ at $\alpha = 1\%$, the null hypothesis $H_0 : \alpha_M = 0$ is rejected. Thus, the intercept parameter of the male linear model is significant. This is to be expected since the intercept parameter is linked to the initial population. If the intercept is zero, this implies that the estimated initial male population is one, which is absurd.

Linear Model for Log Female Population

The total sum of squares, regression sum of squares and the residual sum of squares for the fitted linear model for females are as follows:

$$TSS_F = 13.8729,\ Reg\ SS_F = 13.6647, \text{ and } RSS_F = 0.2082$$

The square of the correlation coefficient is $R_F^2 = Reg\,SS_F / TSS_F = 0.9850$, while the Adjusted $R_F^2 = 0.9840$. That is, 98.4% of the variability between the dependent variable Y_F and the explanatory variable X_F is explained by the fitted linear model for females.

The standard error of the regression and the estimated regression parameters are:

$$\hat{\sigma}_{\varepsilon_F} = \sqrt{RSS_F/(n-2)} = 0.1178,\ SE\left(\hat{\beta}_F\right) = 0.0277, \text{ and } SE\left(\hat{\alpha}_F\right) = 0.0408.$$

The following procedure tests the significance of the regression parameters at level of significance $\alpha = 1\%$ using the two-tailed *t*-test.

Null hypothesis: $H_0 : \beta_F = 0$

t-statistic: $$t_0 = \frac{\hat{\beta}_F - 0}{SE(\hat{\beta}_F)} = \frac{-0.8679}{0.0277} = -31.37$$

Since $|t_0| > t_{\alpha/2} = 3.2860$ at $\alpha = 1\%$, the null hypothesis $H_0 : \beta_F = 0$ is rejected. Hence, the slope parameter of the female linear model is significant.

Null hypothesis: $$H_0 : \alpha_F = 0$$

t-statistic: $$t_0 = \frac{\hat{\alpha}_F - 0}{SE(\hat{\alpha}_F)} = \frac{15.4052}{0.0408} = 377.92$$

Since $t_0 > t_{\alpha/2} = 3.2860$ at $\alpha = 1\%$, the null hypothesis $H_0 : \alpha_F = 0$ is rejected. Thus, the intercept parameter of the female linear model is significant. Again this is to be expected since the intercept parameter is linked to the initial population.

I. Model Diagnostic Checks

(Refer to the Excel regression template for all the required computations)

Residual Plots

The residual plots from the fitted linear regression models Eq-16 and Eq-18 are shown in Figure 6 and Figure 7. Both graphs display increasing residual variances as the explanatory variable increases.

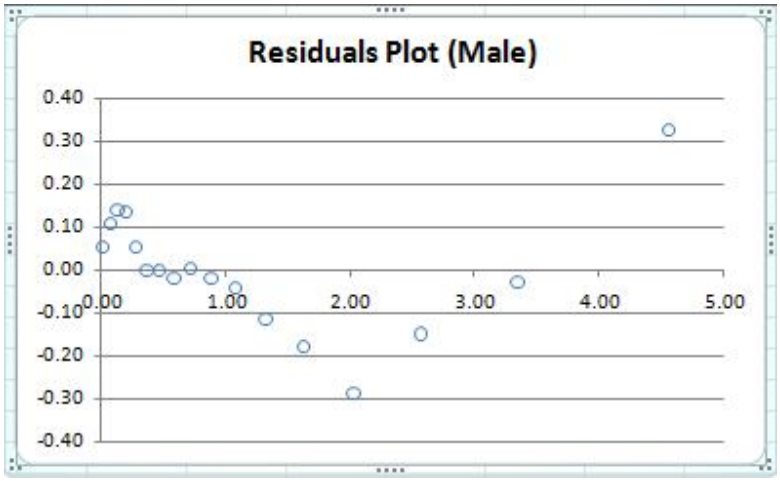

Figure 6. Scatter plot of the residuals from the fitted model Eq-16 (y-axis) and explanatory variable (x-axis)

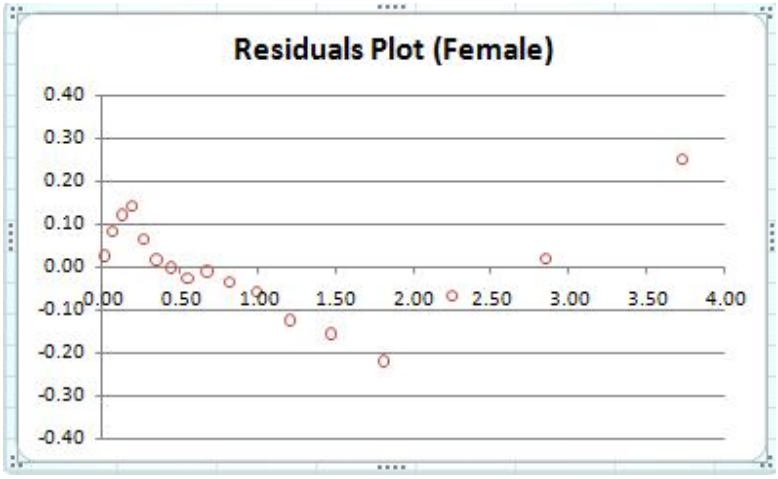

Figure 7. Scatter plot of the residuals from the fitted model Eq-18 (y-axis) and explanatory variable (x-axis)

However, the residual variance is typically "small" in this case. Note that the estimated standard errors of the regression for Eq-16 and Eq-18 are $\hat{\sigma}_{\varepsilon_M} = 0.1450$ and $\hat{\sigma}_{\varepsilon_F} = 0.1178$. For males, the residuals fluctuate between -0.40 and +0.40 while for females, the residuals fluctuate between -0.30 and +0.30, and with the correlation coefficient of more than 98%, the fitted linear models Eq-16 and Eq-18 are quite accurate.

Outliers

The hat values are also calculated to determine if any outlier exist in the data. This is summarized in Table 4 below.

Age Group	Age Variable (x)	X_M	Y_M	Male Hat Values	X_F	Y_F	Female Hat Values
Under 5	2	0.02	15.48	0.1114	0.02	15.41	0.1174
5-9	7	0.08	15.49	0.1066	0.07	15.42	0.1116
10-14	12	0.14	15.47	0.1015	0.13	15.41	0.1055
15-19	17	0.20	15.41	0.0962	0.20	15.38	0.0992
20-24	22	0.28	15.26	0.0906	0.27	15.24	0.0926
25-29	27	0.37	15.13	0.0848	0.35	15.12	0.0858
30-34	32	0.47	15.04	0.0788	0.44	15.02	0.0791
35-39	37	0.59	14.93	0.0729	0.55	14.90	0.0725
40-44	42	0.72	14.83	0.0673	0.68	14.80	0.0665
45-49	47	0.89	14.67	0.0624	0.82	14.65	0.0616
50-54	52	1.08	14.48	0.0593	1.00	14.48	0.0590
55-59	57	1.33	14.20	0.0595	1.21	14.23	0.0603
60-64	62	1.63	13.88	0.0661	1.48	13.97	0.0688
65-69	67	2.03	13.43	0.0855	1.81	13.61	0.0907
70-74	72	2.57	13.11	0.1312	2.25	13.38	0.1381
75-79	77	3.35	12.56	0.2358	2.85	12.95	0.2376
80-84	82	4.56	11.89	0.4900	3.73	12.42	0.4537

For both males and females, the hat values at the old age group "80-84" are typically large at more than 45%. Even for the age group "75-79" the hat values are quite large at more than 23% (approximately half of the hat values for the age group "80-84"). The fitted regression line puts relatively more weights on these age groups, and hence, these age groups tend to "pull" the regression line towards them.

Model Refitting

(Refer to the Excel regression template "refitted version" for all the required computations)

If the age groups "75-79" and "80-84" are removed from the data and the linear models Eq-11 and Eq-15 are refitted on the transformed (truncated) data, the following fitted linear regression models are obtained

$$\hat{Y}_M = 15.5399 - 0.9918 X_M \qquad \text{(Eq-21)}$$

$$(0.0218) \quad (0.0196)$$

$$\hat{Y}_F = 15.4792 - 0.9900 X_F \qquad \text{(Eq-22)}$$

$$(0.0233) \quad (0.0234)$$

The standard errors are shown in parameters. The resulting regression standard errors are reduced by approximately 60% compared to the fitted linear regression models Eq-16 and Eq-18. By the same procedure as hypothesis testing in Model Fitting section, the estimated regression parameters in Eq-21 and Eq-22 are significant at $\alpha = 1\%$ level of significance.

The adjusted square of correlation coefficient for Eq-21 and Eq-22 are 0.9945 and 0.9923, respectively, an almost perfect linear fit. These adjusted squares of correlation coefficient are higher than that obtained from Eq-16 and Eq-18.

The residual plots from the fitted models Eq-21 and Eq-22 are shown in Figure 8 and Figure 9. Compared to Figure 6 and Figure 7, the residual variances are now fairly stable.

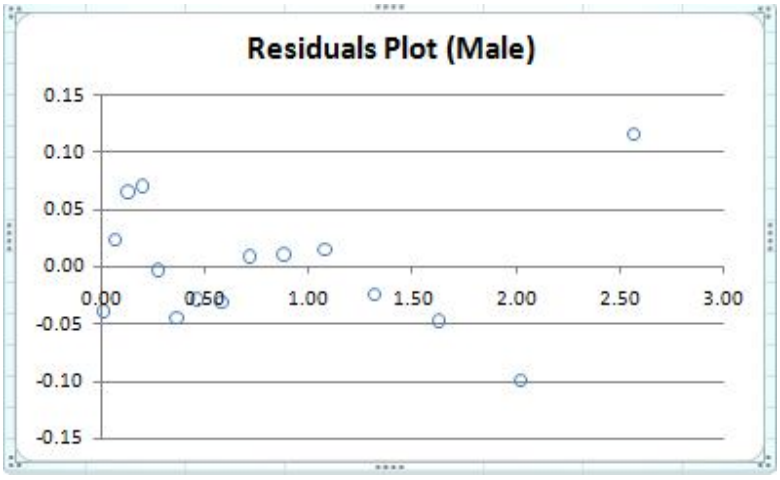

Figure 8. Scatter plot of the residuals from the refitted model Eq-21 (y-axis) and explanatory variable (x-axis)

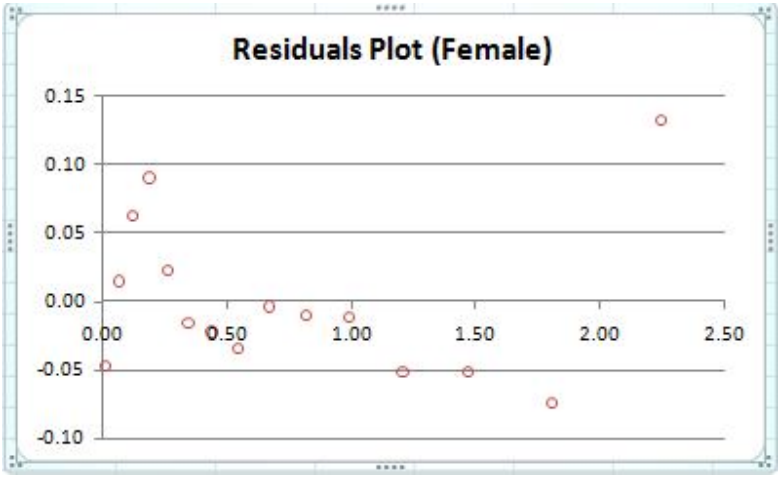

Figure 9. Scatter plot of the residuals from the refitted model Eq-22 (y-axis) and explanatory variable (x-axis)

Combining Eq-21 and Eq-22, and reverting to the "untransformed" models Eq-4 and Eq-5, the final fitted model for the 2010 Philippine Population Data covering ages 0 t0 74 is

$$\hat{P}(x) = 5{,}608{,}869 \exp\left(-\frac{0.9918x}{100-x}\right) + 5{,}278{,}816 \exp\left(-\frac{0.9900x}{104-x}\right) \qquad \text{(Eq-23)}$$

J. Model Interpretation and Forecasting

In the fitted population model Eq-23, the scale parameters are very close to 1. This implies that rate of decrease of the male population approximately the ratio of the current age x to the remaining $100-x$ years of life. Similarly, for females, the rate of decrease in the population is approximately the ratio of current age x to the remaining $104-x$ years of life.

Since the data were truncated at age groups "75-79" and "80-84", the model Eq-23 can be used to forecast the population at these age groups. The following table summarizes the actual and forecasted populations.

Refitted Model Forecast		Forecast	Forecast	Actual	Actual	Error	Error
Age Group	Age (x)	Male	Female	Male	Female	Male	Female
75-79	77	202,727	313,604	285,693	420,284	82,966	106,680
80-84	82	61,196	131,824	145,686	247,701	84,490	115,877

For the same age groups, the following tables show the forecasts from the model Eq-20.

Fitted Model Forecast		Forecast	Forecast	Actual	Actual	Error	Error
Age Group	Age (x)	Male	Female	Male	Female	Male	Female
75-79	77	293,954	412,568	285,693	420,284	-8,261	7,716
80-84	82	105,014	192,991	145,686	247,701	40,672	54,710

Notice that the error of forecasts is smaller in Eq-20 than in Eq-23. This is because age groups "75-79" and "80-84" are included in the fitted model Eq-20 as part of the data. From the analysis of hat values, these age groups are excluded from the model Eq-23. Since these age groups tend to pull the fitted regression lines towards them, excluding these

data from the model "eliminates the pulling effect". Therefore, the forecast values from Eq-23 will not be as close to the actual values as that of Eq-20.

The opposite scenario happens when the populations to be forecasted is "within the data set". This is illustrated in the tables below. The same models as used in the tables above are used but at age groups "20-24" and "50-54".

Refitted Model Forecast		Forecast	Forecast	Actual	Actual	Error	Error
Age Group	Age (x)	Male	Female	Male	Female	Male	Female
20-24	22	4,240,247	4,047,474	4,229,958	4,140,440	-10,289	92,966
50-54	52	1,915,428	1,961,491	1,945,258	1,938,372	29,830	-23,119

Fitted Model Forecast		Forecast	Forecast	Actual	Actual	Error	Error
Age Group	Age (x)	Male	Female	Male	Female	Male	Female
20-24	22	4,009,183	3,883,961	4,229,958	4,140,440	220,775	256,479
50-54	52	2,025,193	2,058,184	1,945,258	1,938,372	-79,935	-119,812

This shows that more accurate models can be used to predict intermediate values within the data set than to extrapolate values beyond the data.