CHAPTER IV
NOVEL USE OF SAS PROGRAMMING IN DEMOGRAPHICS DATA ANALYSIS

4.1 INTRODUCTION AND OUTLINE

It is intended to discuss in some detail about Statistical computing Methodology. It covers use of SAS programming in Demographic Analysis in this Chapter.

4.2 REGRESSION ANALYSIS

Have a clear notion of what you can and cannot do with regression analysis.

- Conceptualization
  - A Path Model of a Regression Analysis

Path Diagram of A Linear Regression Analysis

- Hypothesis Testing
  - For example: Hypothesis 1: X is statistically significantly related to Y.
– The relationship is positive (as X increases, Y increases) or negative (as X decreases, Y increases).
– The magnitude of the relationship is small, medium, or large. If the magnitude is small, then a unit change in x is associated with a small change in Y.

**Plotting the Data: Testing for Functional form**

1. The relationship between the Y and each X needs to be examined.
2. One needs to ascertain whether it is linear or not
3. If it is not linear, can it be transformed into linearity?
4. If so, then that may be the easiest recourse.
5. If not, then perhaps linear regression is not appropriate

**Graphical Exploration**

We examine the plot to look for functional form,

Outlier patterns and other peculiarities.

The generation of the data is done with a do loop:
```plaintext
data plot1;
do x = 1 to 30;
y = x**3;
output;
end;
proc plot;
plot y*x; run;
```

**Graphical Plot**

- The output from this plot appears in figure 1:

![Figure 1](image)

**Testing Functional Form with Curve Fitting**

1. Curve Fitting
2. What is done here is that the data are subjected to a number of tests of functional form. Application of the data to such a process presumes either a linear relationship or a relationship that can be transformed into a linear one.

1. From a regression analysis of the relationship, an r square is generated. This is the square of the multiple correlation coefficient between the dependent variable and the independent variables in the model. This r square is the proportion of variance of the dependent variable explained by the functional form. The higher the r square, the closer the functional form is to the actual relationship inherent in the data.

We may fit functional forms

```plaintext
Do x = 1 to 100;
liny  = a + x;
quady = a + x**2;
cubicy = a + x**3;
lnx   = log(x);
invx  = 1/x;
expx  = exp(x);
compound = a*b**x;
power = a*x**b;
sshape = exp( a + x**(-1));
growth = exp(a + b*x);
output;
End;
Proc print;
run;
Then run different models. For example,
Proc Reg;
   model y = quadx;
   run;
Use the transformation that yields the highest R-square
```

Curve Fitting Output & Interpretation
Output and interpretation

The proc print label data = bbb;
Var _model_ p rsq;
Run;

Produces the output in figure 5 on the next page.

While all three forms have significant components, it can be seen that the $R^2$ approaches 1.00 only with the cubic and power functional forms. The regression coefficient for this curve equals 1.00 and we have identified the functional form of the regression analysis. The cubic functional form is the third power of the function, so it is not surprising that the coefficient of determination ($R^2$) for that will be 1.00 also. The nice aspect of curve fitting is that it provides a clear objective criterion against which to assess functional form.

With the aid of this tool, one may obtain a better idea as to which transformation to apply in order to render a nonlinear relationship amenable to linear regression, in the event of apparent nonlinearity.

Linear Regression analysis
A. The General Linear Model
B. Derivation of the SS and Anova
C. Derivation of the coefficients: a and b
   1. Bivariate Case
   2. Multivariate Case: $a$, $b_1$ and $b_2$
D. The significance tests
E. The Prediction interval and its derivation
F. The Assumptions of Regression Analysis
G. Testing those assumptions
H. Robustness of the Model in face of their violation
I. Fixes:
   1. Weighted Least Squares estimation for heteroskedasticity
   2. Autoregression for autocorrelation
   3. Nonlinear regression for nonlinearity

The General Linear Model
A. The General Linear Model
   Includes regression, anova, and anova. The main difference is the levels of measurement of the independent variables:
   1. Regression: IVs are continuous
   2. Anova: IVs are discrete
   3. Ancova: IVs are both discrete and continuous

B. Decomposition of the Sums of squares
   If we look at the graph of the regression line, we may geometrically represent the error, the regression effect, and the total effect.

   The graphical representation below depicts the sums of squares decomposition of the equation:

   \[ SS_{total} = SS_{regression} + SS_{error} \]

   If we divide both sides of the equation by the respective df, we obtain the MS:

   \[ \frac{SS_{total}}{n-1} = \frac{SS_{regression}}{k} + \frac{SS_{error}}{n-k-1} \]
   Where \( n \) = sample size and \( k \) = # independent vars in model

   \[ MS_{total} = MS_{regression} + MS_{error} \]

   Since \( MS = \text{variance} \)
   Variance total = Regression Variance + error variance
   We divide both sides by the total variance to obtain

   \[ 1 = \frac{R^2}{\text{sum of squares regression}} + \frac{R^2}{\text{sum of squares error}} \]

   \[ F = \frac{\text{Regression Variance}}{\text{Error Variance}} \]

Decomposition of the sum of squares
Decomposition of the sum of squares

\[ Y - \bar{Y} = \hat{Y} - Y + \hat{Y} - \bar{Y} \]

**total effect** = error effects + regression (model) effect

\[ Y_i - \bar{Y} = \hat{Y}_i - Y_i + \hat{Y}_i - \bar{Y} \] **per case** \( i \)

\[ (Y_i - \bar{Y})^2 = (\hat{Y}_i - Y_i)^2 + (\hat{Y}_i - \bar{Y})^2 \] **per case** \( i \)

\[
\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \] **for data set**

**F test for significance and R-square for magnitude of effect**

\[ R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{k} \]

\[ R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{\frac{n - k - 1}{k}} \]

**F test for model significance**

= Model Var/Error Var
Derivation of the Intercept

\[ y = a + bx + e \]
\[ e = y - a - bx \]
\[ \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_i - b \sum_{i=1}^{n} x_i \]

Because by definition \[ \sum_{i=1}^{n} e_i = 0 \]
\[ 0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_i - b \sum_{i=1}^{n} x_i \]
\[ \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i \]
\[ na = \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i \]
\[ a = \bar{y} - b \bar{x} \]

Derivation of the Regression Coefficient
Given: \( y_i = a + b x_i + e_i \)
\( e_i = y_i - a - b x_i \)
\[ \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a - b x_i) \]
\[ \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - b x_i)^2 \]
\[ \frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b} = 2x_i \sum_{i=1}^{n} (y_i) - 2b \sum_{i=1}^{n} x_i x_i \]
\( b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \)

The Multiple Regression Equation

- We proceed to the derivation of its components:
  - The intercept: a
  - The regression parameters, b1 and b2

\[ Y_i = a + b_1 x_1 + b_2 x_2 + e_i \]

- If we recall that the formula for the correlation coefficient can be expressed as follows:
Significance Tests for the Regression Coefficients

1. We find the significance of the parameter estimates by using the F or t test.

1. The R-square is the proportion of variance explained.
2. Adjusted R-square
   \[ = 1 - (1 - \text{R-square}) \frac{(n-1)}{(n-p-1)} \]

F and T tests for significance for overall model

\[ F = \frac{\text{Model variance}}{\text{error variance}} = \frac{R^2 / p}{(1-R^2)/(n-p-1)} \]

where
\[ p = \text{number of parameters} \]
\[ n = \text{sample size} \]
The Multiple Regression Equation

We proceed to the derivation of its components:
- The intercept: $a$
- The regression parameters, $b_1$ and $b_2$

\[ Y_i = a + b_1 x_1 + b_2 x_2 + e_i \]
The Multiple Regression Formula

Pearson’s Correlation Coefficient

\[ r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} (x_i^2) \sum_{i=1}^{n} (y_i^2)}} \]

where

\[ x = x_i - \bar{x} \]
\[ y = y_i - \bar{y} \]

\[ b_j = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x^2} \]

from which it can be seen that the regression coefficient \( b \) is a function of \( r \).

\[ b_j = r \times \frac{sd_y}{sd_x} \]

Substitute the partial \( r \) for \( r \)

\[ \beta_{y_1,x_2} = \frac{r_{y_1,x_2} - r_{y_1,x_2} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \times \frac{sd_y}{sd_x} \]  \hspace{1cm} (6)

\[ \beta_{y_2,x_1} = \frac{r_{y_2,x_1} - r_{y_2,x_1} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \times \frac{sd_y}{sd_x} \]  \hspace{1cm} (7)

It is also easy to extend the bivariate intercept to the multivariate case as follows.

\[ a = \bar{y} - b_1 \bar{x_1} - b_2 \bar{x_2} \]  \hspace{1cm} (8)
Significance Tests for the Regression Coefficients

1. We find the significance of the parameter estimates by using the F or t test.

2. The $R^2$ is the proportion of variance explained.

3. Adjusted $R^2$

$$= 1 - \frac{(1-R^2) (n-1)}{(n-p-1)}$$

F and T tests for significance for overall model

$$F = \frac{Model\ variance}{error\ variance}$$

$$= \frac{R^2 / p}{(1-R^2)/(n-p-1)}$$

where

$p =$ number of parameters

$n =$ sample size

$$t = \sqrt{F}$$

$$= \sqrt{\frac{(n-2) \cdot p^2}{1-r^2}}$$
Significance tests

T-tests for statistical significance

\[ t = \frac{\alpha - \theta}{se_a} \]

\[ t = \frac{b - \theta}{se_b} \]

Standard Error of intercept

\[ SE_a = \left[ \frac{\sqrt{(Y - \bar{Y})^2}}{n-2} \right] \times \left[ \frac{1}{n} + \frac{\sum x_i^2}{(n-1)\sum (x_i - \bar{x})^2} \right] \]

Standard error of regression coefficient

\[ SE_b = \frac{\hat{\sigma}}{\sum x^2} \]

where \( \hat{\sigma} = \text{std dev of residual} \)

\[ \hat{\sigma}^2 = \frac{\sum e^2}{n-1} \]
SAS Regression Command Syntax

Figure 9 Command Syntax for Regression Analysis of Education on possible influences
SAS Regression syntax

Proc reg simple data=regdat;
model y = x1 x2 x3 /spec dw corrb collin r dffits influence;
output out=resdat p=pred
r=resid rstudent=rstud;
Data check;
set resdat;
Proc univariate normal plot;
var resid;
Title 'Check of Normality of Residuals';
Run;
Proc Freq data=resdat; tables rstud;
Title 'Check for Outliers';
run;
Proc ARIMA data=resdat;
identify var=resid;
Title 'Check of Autocorrelation of the errors';
Run;

SAS Regression Output & Interpretation

![Image of SAS output]

Figure 1. Sample statistics of variables included in the Regression model.
More Simple Statistics

Figure 1 More Simple Statistics of Variables in Model

Parameter Estimates

Figure 1 Parameter Estimates, Standard Errors, T-tests, significance levels, Type-I and Type-II Sums of Squares, Squared Semi-partial Correlations of Type I and II, and the variable labels for variables in the Background Model.
Assumptions of the Linear Regression Model

1. Linear Functional form
2. Fixed independent variables
3. Independent observations
4. Representative sample and proper specification of the model (no omitted variables)
5. Normality of the residuals or errors
6. Equality of variance of the errors (homogeneity of residual variance)
7. No multi-collinearity
8. No autocorrelation of the errors
9. No outlier distortion

Explanation of the Assumptions

1. 1. Linear Functional form
   1. Does not detect curvilinear relationships
2. Independent observations
   1. Representative samples
   2. Autocorrelation inflates the t and r and f statistics and warps the significance tests
3. Normality of the residuals
   1. Permits proper significance testing
4. Equality of variance
   1. Heteroskedasticity precludes generalization and external validity
   2. This also warps the significance tests
5. Multi-collinearity prevents proper parameter estimation. It may also preclude computation of the parameter estimates completely if it is serious enough.
6. Outlier distortion may bias the results: If outliers have high influence and the sample is not large enough, then they may serious bias the parameter estimates
Diagnostic Tests for the Regression Assumptions

1. Linearity tests: Regression curve fitting
   1. No level shifts: One regime
2. Independence of observations: Runs test
3. Normality of the residuals: Shapiro-Wilks or Kolmogorov-Smirnov Test
4. Homogeneity of variance if the residuals: White’s General Specification test
5. No autocorrelation of residuals: Durbin Watson or ACF or PACF of residuals
6. Multicollinearity: Correlation matrix of residuals. Condition index or condition number
7. No serious outlier influence: tests of additive outliers:
   1. Plot residuals and look for high leverage of residuals
   2. Lists of Standardized residuals
   3. Lists of Studentized residuals
   4. Cook’s distance or leverage statistics

Explanation of Diagnostics

1. Plots show linearity or nonlinearity of relationship
2. Correlation matrix shows whether the independent variables are collinear and correlated.
3. Representative sample is done with probability sampling

Tests for Normality of the residuals. The residuals are saved and then subjected to either of:
Kolmogorov-Smirnov Test: Tests the limit of the theoretical cumulative normal distribution against your residual distribution.
Shapiro-Wilks Test

Proc reg;
   model y = x1 x2;
   Output out=resdat r=resid p=pred;
Data check;
   set resdat;
Proc univariate normal plot; var resid;
Title ‘Test of Normality of Residuals’;
Run;
Test for Homogeneity of variance

White’s General Test
\[ e_i^2 = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2 \]

Proc reg;
  model y = x1 x2/spec;
Run;

Chow test

\[ F_{\text{chow}} = \frac{(R_{wr}^2 - R_r^2) / (p_{wr} - p_r)}{(1 - R_{wr}^2) / (n - p_{wr})} \]

Weighted Least squares

A fix for heteroskedasticity is Weighted least squares or WLS.
There are two ways to do this.
If \( x \) is proportional to the variance in \( y \), then form the weight by \( wt = 1/y^{**2} \);
Proc reg;
  weight wt;
Model y = x;
Run;
Collinearity Diagnostics

**Tolerance** = \( 1 - R^2 \)
small tolerances imply problems

**Variance Inflation Factor (VIF)**

\[ VIF = \frac{1}{\text{Tolerance}} \]
Small intercorrelations among indep vars
means VIF ≈ 1

\( VIF > 10 \) signifies problems

More Collinearity Diagnostics

Watch for eigenvalues much greater than 1
condition numbers = maximum eigenvalue/minimum eigenvalue.
If condition numbers are between 100 and 1000, there is moderate to strong collinearity

condition index = \( \sqrt{k} \)
where k = condition number

If Condition index > 30 then there is strong collinearity
Collinearity Diagnostics

![Collinearity Diagnostics](image)

**Figure 1** Collinearity Diagnostics: Eigenvalues, Condition Indices, etc.

Outlier Diagnostics

1. Residuals.
   1. The predicted value minus the actual value. This is otherwise known as the error.
2. Studentized Residuals
   1. the residuals divided by their standard errors without the ith observation
3. Leverage, called the Hat diagram
   1. This is the measure of influence of each observation
4. Cook’s Distance:
   1. the change in the statistics that results from deleting the observation. Watch this if it is much greater than 1.0.
Outlier detection

- Outlier detection involves the determination whether the residual (error = predicted – actual) is an extreme negative or positive value.
- We may plot the residual versus the fitted plot to determine which errors are large, after running the regression.
- The command syntax was already demonstrated with the graph on page 16: rvfplot, border yline(0)

Create Standardized Residuals

- A standardized residual is one divided by its standard deviation.

\[
\text{resid}\,_{\text{standardized}} = \frac{\hat{y}_i - y_i}{s}
\]

where \( s = \text{std dev of residuals} \)
Limits of Standardized Residuals

- If the standardized residuals have values in excess of 3.5 and -3.5, they are outliers.
- If the absolute values are less than 3.5, as these are, then there are no outliers.
- While outliers by themselves only distort mean prediction when the sample size is small enough, it is important to gauge the influence of outliers.
- Suppose we had a different data set with two outliers.
- We tabulate the standardized residuals and obtain the following output:

Studentized Residuals

- Alternatively, we could form studentized residuals. These are distributed as a $t$ distribution with $df=n-p-1$, though they are not quite independent. Therefore, we can approximately determine if they are statistically significant or not.
- Belsley et al. (1980) recommended the use of studentized residuals.

$$e_i^s = \frac{e_i}{\sqrt{s^2_{(i)} (1 - h_i)}}$$

where

- $e_i^s = \text{studentized residual}$
- $s_{(i)} = \text{standard deviation where ith obs is deleted}$
- $h_i = \text{leverage statistic}$

These are useful in estimating the statistical significance of a particular observation, of which a dummy variable indicator is formed. The $t$ value of the studentized residual will indicate whether or not that observation is a significant outlier.

The command to generate studentized residuals, called `rstudent` is:

```
predict rstudt, rstudent
```
Graphical Exploration of Outlier Influence

One can plot the leverage against the standardized residual to see if the outlier is problematic.

Alternatives to Violations of Assumptions

- **Nonlinearity**: Transform to linearity if there is nonlinearity or run a nonlinear regression.
- **Nonnormality**: Run a least absolute deviations regression or a median regression.
- **Heteroskedasticity**: Weighted least squares regression or white estimator. One can use Proc Robustreg to obtain downweighted outlier effect in the estimation.
- **Autocorrelation**: Newey-West estimators or autoregression model.
- **Multicollinearity**: Components regression or ridge regression or proxy variables.
The Interaction model

\[ Y_i = a + b_1x_1 + b_2x_2 + b_{ij}x_1^*x_2 + \epsilon_i \]

- The product vector \( x_1^*x_2 \) is an interaction term.
- This is the joint effect over and above the main effects (\( x_1 \) and \( x_2 \)).
- The main effects must be in the model for the interaction to be properly specified, regardless of whether the main effects remain statistically significant.

Path Diagram of an Interaction model

Interaction Analysis

\( Y = K + aX_1 + BX_2 + CX_1^*X_2 \)
SAS Regression Interaction
Model Syntax

```sas
data one;
input y x1 x2 x3 x4;
lincrossprodx1x2=x1*x2;
x1sq=x1*x1;
time=1;
datalines;
112 113 114 39 10
322 230 310 43 23
323 340 250 33 33
112 122 125 144 45
99 100 89 55 34
14 13 10 249 40
40 34 98 39 30
30 32 40 40
90 93 50 50
89 90 60 44
120 130 43 100 34
44 432 430 20 44
proc print;
run;
proc reg;
model y= x1 x2 lincrossprodx1x2/r collin stb spec influence;
output out=resdat r=resid p=pred stdr=rstd student=rstud cookd=cooksd;
run;
```

SAS Regression Diagnostic
syntax

```sas
data outck;
set resdat;
degfree=7;
if cooksd > 4/7; /* if cd > p/(n-p-1)
where p = # parms, n=sample size */
proc freq; tables rstud;
title 'Outlier Indicator';
run;
axis1 label=(a=90 'Cooks D Influence Stat');
proc gplot data=resdat;
plot cooksd * rstud/vaxis=axis1;
title 'Leverage and the Outlier';
run;
```
Regression Model test of linear interaction

The SAS Procedure
Model: MODEL1
Dependent Variable: y

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>72113</td>
<td>36556.5</td>
<td>182.79</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>72113</td>
<td>36556.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>29</td>
<td>144226</td>
<td>5130.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square 0.9953
Dependent Mean 141.5533
Adj R-Sq 0.9953

Parameter Estimates

| Variable     | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| Standardized Estimate |
|--------------|----|--------------------|----------------|---------|------|-----------------------|
| Intercept    | 1  | -0.87122           | 14.03264       | -0.06 | 0.9504 | 0                    |
| x1           | 1  | -0.26483           | 0.18801        | -1.43  | 0.1692 | 0.2098 |
| x1sq         | 1  | 0.01116            | 0.00872        | 1.27   | 0.2131 | 0.2486 |
| Tours_predicted | 1  | 0.000005464 | 0.0000000304 | -1.73  | 0.0906 | 0.1053 |

SAS Polynomial Regression Syntax

```sas
proc reg data=one;
   model y= x1 x1sq;
   title 'The Polynomial Regression';
run;
```
Model Building Strategies

1. Specific to General: Cohen and Cohen
2. General to Specific: Hendry and Richard
   Loess, Splines for mode fitting, data reduction, missing data imputation, validation & simplification.

Goodness of Fit indices

- \( R^2 = 1 - \frac{SSE}{SST} \)
  - Explained Variance divided by total variance
- \( AIC = -2LL + 2p \)
  \( SC = -2LL + n\log(p) \)

- Nested models can be compared on the basis of these criteria to determine which model is best
Robust Regression

1. This procedure permits 4 types of robust regression: M, least trimmed squares, S, and MM estimation
2. These means down-weight the outliers.
3. M (median absolute deviation) estimation used by Huber is performed with
4. proc robustreg data=stack; model y = x1 x2 x3 / diagnostics leverage; id x1; test x3; run;
5. Estimation is done by IRLS.

4.3 EXAMPLE OF A LINEAR REGRESSION MODEL IN SAS SOFTWARE

Norman Ryder, a professor emeritus of Sociology at Princeton University conducted pioneering studies of fertility in the United States. The work of Norman Ryder and Charles Westoff served as the model for continued global fertility studies. Their studies were incorporated by the National Center for Health Statistics in its National Survey of Family Growth. Ryder and Westoff also prepared the core questionnaire for the World Fertility Survey of the International Statistical Institute, to which Ryder served as consultant from 1972 to 1984. Fertility surveys were conducted in some 40 developing countries through that program, which often is described as one of the 20th century's most important international undertakings in demographic data collection and analysis.

In OECD countries, there is a strong correlation between unemployment rate and fertility rate as revealed by such surveys.

4.3.1 DETERMINANTS OF FERTILITY RATES
Most analyses of childbearing decisions have their root in the economic model pioneered by Becker (1960) and Leibenstein (1957), where demand for children is a function of their costs and of individuals’ preferences, for a given income. Underlying this model is the notion that children are a special type of capital good, *i.e.* a long-lived asset that produces a flow of services that enter the utility function of parents. This model provides a framework for analyzing the effects on fertility rates of various economic and social factors, including literacy rates.

As educated among women has increased, more women are going to college from backgrounds that previously would have precluded their attendance and completion. This affords us the opportunity and motivation to look at the effects of education on fertility across a range of social backgrounds and levels of early achievement.

Disaggregating the effects of education level and literacy rate, we find that the fertility-decreasing education effect is concentrated among women from comparatively disadvantaged social backgrounds and low levels of early achievement. Also there is a difference between rural and urban education levels (Brand and Davis).

**Table 4.3.1: MODELING TOTAL FERTILITY RATE for 29 Indian States & 6 Union Territories**

*(Based on data from National Rural Health Mission (NRHM) & SRS – Sample Registration System/Govt. of India)*

Model: Linear Regression Model

**Dependent Variable: Total Fertility Rate (TFR)**

<table>
<thead>
<tr>
<th>Number of Observations Read</th>
<th>34</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations Used</td>
<td>27</td>
</tr>
<tr>
<td>Number of Observations with Missing Values</td>
<td>7</td>
</tr>
</tbody>
</table>

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Mean</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
</table>
### Squares Square

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>10.89690</td>
<td>1.81615</td>
<td>45.06</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>0.80606</td>
<td>0.04030</td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>26</td>
<td>11.70296</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Root MSE</td>
<td>0.20076</td>
<td>R-Square</td>
<td>0.9311</td>
<td></td>
</tr>
<tr>
<td>Dependent Mean</td>
<td>2.56296</td>
<td>Adj R-Sq</td>
<td>0.9105</td>
<td></td>
</tr>
<tr>
<td>Coeff Var</td>
<td>7.83299</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Parameter Estimates

| Variable                                  | DF | Parameter Estimate | Parameter Standard Error | t | Pr > |t| |
|-------------------------------------------|----|--------------------|--------------------------|---|------|---|
| Intercept                                 | 1  | 0.35650            | 0.38688                  | 0.92 | 0.3678 | |
| Mother's with no education                | 1  | 0.56879            | 0.05457                  | 10.42 | <.0001 | |
| Mother's with >10 years of educa          | 1  | 0.37078            | 0.14038                  | 2.64 | 0.0157 | |
| GDP per Capita (USD)                      | 1  | 0.00010511         | 0.00007394               | -1.42 | 0.1706 | |
| Sex Ratio                                 | 1  | 0.00048590         | 0.00025033               | 1.94 | 0.0665 | |
| Population below poverty line             | 1  | 0.00076415         | 0.00428                  | -0.18 | 0.8601 | |
| Female Literacy Rate                      | 1  | -0.01261           | 0.00387                  | -3.26 | 0.0039 | |

4.4 **CONCLUSION BASED ON KEY INDICATORS:**

In the SAS Regression shown above, the prediction equation is
TFR = 0.35650 + 0.56879 (Mothers with/No education) + 0.37078 (Mothers w > 10 years of education) - 0.00005 (Sex Ratio) – 0.0002 (GDP per capita)-0.01261 (Female literacy Rate) - 0.0008 (Population below Poverty Line)

The results are telling us that TFR is predicted to increase by 0.56879 when the Mother w/no education variable goes up by one, increase by 0.37078 when the Mothers w>10 years education variable increase by one. Decrease by 0.00005 when sex ratio goes up by one and decrease by 0.00021 when GDP per capita increase by one. The TFR is predicted to be 0.35 when all the variables are zero.

The t-statistic is the coefficients divided by its standard error, which is an estimate of the standard deviation of the coefficient. It can be thought of as a measure of the precision with which the regression coefficient is measured. If a coefficient is large compared to its standard error, then it is probably different from zero. So we look at the p-value for significance level.

Mother w/no education is the most significant variable and is significant at 99% confidence interval. Mothers w>10 years of education is the next important variable followed by GDP per capita which has a negative relationship with TFR. In other words if income increase, fertility rate goes down as working mothers may not have time or energy to bear children.

The R-square of the regression is the fraction of the variation in the dependent variable that is predicted by the independent variables. In this regression, the R-square is 0.88 (high correlation).

Higher education influences women’s career aspirations, labor market involvement and experiences, familial roles, and fertility.

Educated women delay the onset of childbearing and have fewer children overall compared with less-educated women.
Figure 4.4.1 Distribution of Residuals for Total Fertility Rate

Linear

Figure 4.4.2: Residual by Predicted for Total Fertility Rate
Figure 4.4.3: R Student by Predicted for Total Fertility Rate

Figure 4.4.4: Observed (Data) by Predicted for Total Fertility Rate
**Figure 4.4.5** : Cook’s D for Total Fertility Rate

**Figure 4.4.6** : Outlier and Leverage Diagnostics for Total Fertility Rate
Figure 4.4.7: Q-Q Plot of Residuals for Total Fertility Rate

Figure 4.4.8: Residual – Fit Spread Plot for Total Fertility Rate
Figure 4.4.9: Residuals for Total Fertility Rate

Figure 4.4.10: Influence Diagnostics for Total Fertility Rate
4.3.2 REGRESSION MODEL DIAGNOSTICS
Several kinds of regression diagnostics are available in SAS: co-linearity, influential observations, and residual plots. We'll start with looking at residual plots. Studentized residuals are the ordinary residuals divided by their standard errors where all calculations are done after dropping each observation one at a time. I.e., for the first Studentized residual, the first observation is dropped and parameter estimates obtained. A residual is then obtained for the first observation using these parameter estimates. An observation with an absolute Studentized value greater than 2 or 3 is considered an outlier.

4.4 SUMMARY

Educated women delay the onset of childbearing and have fewer children overall compared with less-educated women. In the regression analysis on TFR, the literacy rate is a significant predictor of explaining the fertility in any population. An educated mother is an informed mother who is aware of nutrition, family planning and other areas of public health. For example in Kenya, women bear 5 children compared to China where the one-child policy is in effect. In prosperous China, the economic boom is equally distributed among the population and the government policy of housing etc. has raised the standard of living for all. India is yet to get to that level as its policies are not inclusive of the general population, with wide disparities in income and gender. In India, the lowest TFR of 1.8 is in Kerala where the literacy rate is very high comparable to Western countries.