# *Chapter 5*

# Linear Regression Models

Aim: Upon completing the chapter, the learner should be able to use *simple, multiple,* and *polynomial linear regression models* for estimating the expected values of a continuous random variable.

## 5.1 Introduction

A simple linear regression model (SLRM) is a statistical technique that attempts to model the relationship between two variables. One of these variables is the main outcome of interest and is a quantitative random variable, usually denoted with the letter $Y$ and called the *response* or *dependent* variable. The second one can also be quantitative and is used to explain the behavior of the expected values of $Y$; it is usually denoted with the letter $X$ and is called the *predictor, explanatory,* or *independent* variable. The relationship between these variables, when $X$ is a quantitative variable, is established using the following expression:

$$\mu_{y_i|x_i} = \beta_0 + \beta_1 * x_i$$

where:

- $\mu_{y_i|x_i}$ defines the expected value of the random variable $Y$ given the predictor variable $X$ for the $i$th subject
- $\beta_1$ is a constant parameter associated with the predictor variable $X$; it is known as the *slope* of the regression line and indicates the change in the expected value of $Y$ per unit of change in $X$
- $\beta_0$ is a constant parameter that indicates the expected value of $Y$ when $x_i = 0$; it is known as the *intercept* of the regression line

A simple regression model can also be expressed with the following formula:

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

where:

$y_i$ indicates the response or dependent variables for the $i$th subject

$e_i$ denotes the *residual,* which is the difference between the observed values in $Y_i$ and the expected value under the model $\beta_0 + \beta_1 * x_i$ for the $i$th subject, as follows:

$$e_i = (y_i - \beta_0 + \beta_1 * x_i)$$

## 5.2 Model Assumptions

The procedure to estimate the regression coefficients of an SLRM is performed based on the following assumptions:

1. The response variable is a quantitative random variable that follows a normal distribution with an expected value of $\beta_0 + \beta_1 * x_i$ and a variance of $\sigma^2_{Y|X}$.

   $$Y \sim N\left(\beta_0 + \beta_1 * x_i, \sigma^2_{Y|X}\right)$$

   The expected value of the random variable $Y$ is a straight-line function of $X$.
2. There is independence between the response variable values.
3. The independent or predictor variable is a quantitative variable, not necessarily a random one.
4. The $\beta_i$ coefficients should not be affected by any power, other than the unit, or by any trigonometric function.
5. The expected value of the residuals is zero, that is,

   $$E\left(e_i\right) = 0 \quad \text{for every value of "}i\text{"}$$

6. The variance of the residuals is constant and is equal to the variance of the response variable under the SLRM, that is,

   $$\text{var}\left(e_i\right) = \sigma^2_{Y/X}$$

   This variance is constant across the range of values of $X$ (*Homoscedasticity* property).
7. There is no correlation between the residuals, in that

   $$E\left(e_i, e_j\right) = 0 \quad \text{for all, } i \neq j$$

   The residual associated with a subject does not affect the residual of another subject.

8. The probability distribution of the residuals is normal, as can be determined using the following:

$$e_i \sim N\left(0, \sigma^2_{Y/X}\right)$$

## 5.3 Parameter Estimation

Several methods are available for estimating the beta coefficients for the linear regression model. In classical statistics, the method of *least squares* is used. This method chooses the coefficients that minimize the following residual sum of squares:

$$S = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

where:

$y_i$ indicates the observed value of the $y$ variable for the $i$th subject

$\hat{y}_i$ indicates the estimated expected value of $Y$ under the model for the $i$th subject using a specific combination of the estimated beta coefficients, as follows:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

To reach the coefficients that minimize $S$, the mathematical method of optimization is used, which equates the first derivative of $S$ to 0 (Draper and Smith, 1998). As a consequence, the resulting equation $\left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$ minimizes the distance between the fitted values ( $\hat{y}_i$ ) and the observed values ($y_i$).

## 5.4 Hypothesis Testing

To assess the statistical significance of the changes in the expected value of $Y$ per unit of change in $X$, any one of several methods can be used (Bingham and Fry, 2010). The null hypothesis of these methods states that the coefficient associated with the predictor variable is zero ($H_0$: $\beta_1 = 0$). One of the methods for testing this hypothesis is the $t$-test, which is expressed in the following way:

$$t = \frac{\hat{\beta}_1}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} \sim t_{n-2|H_0}$$

in which $\hat{\beta}_1$ indicates the estimated $\beta_1$ , $t_{n-2}$ is the *t-probability distribution* with $n - 2$ degrees of freedom under the null hypothesis, and $\text{Var}\left(\hat{\beta}_1\right)$ is the variance of $\beta_1$. To compute the $P$-value, it is assumed that this formula follows the $t$-probability distribution under the null hypothesis assumption.

Another method is called *analysis of variance* (ANOVA), which decomposes the variance of *Y*, as follows:

$$\text{TSS} = \text{SSR} + \text{SSE}$$

where:

- $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ indicates the variation of *Y* around the overall mean of *Y*; this variation is known as the *total sum of squares*
- $\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ indicates the variation due to the *model,* the explained variation between the expected value under the model and the overall mean of *Y*; this variation is known as *the sum of squares* (SS) due to the regression model
- $\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y})^2$ indicates the overall variation *within* the model and the unexplained variation within the residuals; this variation is known as the *residual sum of squares* or *error sum of squares*

SS divided by their degrees of freedom are the *mean squares* (MS). The following ANOVA table summarizes the sources of variation in the data, SS due to the source, degrees of freedom in the source, MS due to the source, and the expected value of the MS (Draper and Smith, 1998):

| *Source of Variation* | *Sum of Squares* (*SS*) | *Degrees of Freedom* (*df*) | *Mean Squares* (*MS* = *SS*/*df*) | *Expected Value of MS* |
|---|---|---|---|---|
| Regression | SSR | 1 | SSR/1 | $\sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$ |
| Residual (error) | SSE | $n - 2$ | SSE/($n - 2$) | $\sigma^2$ |
| Total | TSS | $n - 1$ | | |

Under the null hypotheses, the ratio $\left\{\left[\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2\right] / \sigma^2\right\} = 1$. To determine how far this ratio should be away from 1, once a set data are collected and the parameters of the linear model $(\sigma^2, \beta_i)$ are estimated, a *P*-value is computed using the *F*-Fisher probability distribution with 1 and $n - 2$ degrees of freedom.

## 5.5 Coefficient of Determination

The coefficient of determination $R^2$ is a measure of the goodness of fit of the model and is defined using the following expression:

$$R^2 = \frac{\text{SSR}}{\text{TSS}} \times 100\%$$

where the range of values of $R^2$ is between 0% and 100%. This coefficient determines the percentage of variation of the variable *Y* explained by the model. Another way to calculate $R^2$ is with the following formula:

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}} \times 100\%$$

This coefficient is used as a criterion to compare two or more models; the higher the $R^2$, the better the model fits the data.



## 5.6 Pearson Correlation Coefficient

The Pearson correlation coefficient is an index indicating the direction and strength of the linear association between two continuous random variables. This coefficient is represented by $\rho$; the estimator is represented by $r = \hat{\rho}$. Its values range from –1.0 to 1.0. If $r$ is close to 1.0 or to –1.0, it is said that there is a strong positive (directly proportional) or negative (inversely proportional) linear association, respectively; values close to zero indicate little or no linear association. The mathematical expression of $r$ is the following:

$$r = \frac{\text{SSXY}}{\sqrt{\text{SSX} * \text{TSS}}}$$

where:

$$\text{SSXY} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

and

$$\text{SSX} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

An alternative way to calculate the Pearson correlation coefficient is using the square root of the coefficient of determination, assigning the sign of the $\beta_1$ previously estimated:

$$r = \text{sign}\left(\hat{\beta}_1\right)\sqrt{R^2}$$

To assess whether the Pearson correlation coefficient is different from zero ($H_0$: $\rho = 0$), with data from a random sample of size $n$, the following formula is used (Kleinbaum et al., 2008):

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

To obtain the *P*-value, the *t-distribution* with $n - 2$ degrees of freedom is used. This test is equivalent to the *t*-test for assessing $H_0$: $\beta_1 = 0$, described previously.

## 5.7 Scatter Plot

Prior to performing a linear regression analysis for specific data, it is recommended that a scatter plot of the observed data be done to determine whether a linear trend can be associated with the relationship that exists between the dependent and independent variables. For example, the relationship between the variables *weight* and *height* (from the previous database) can be obtained (Figure 5.1) with the following command:

```
twoway (scatter weikg heimt) (lfit weikg heimt),
xtitle("Height, mt") ytitle("Weight, kg")
```

*Output*

The option *lfit* in the *twoway* command is used to draw a line to describe the linear relationship between two variables using an SLRM: in this case between *height* and *weight* given the observed data.

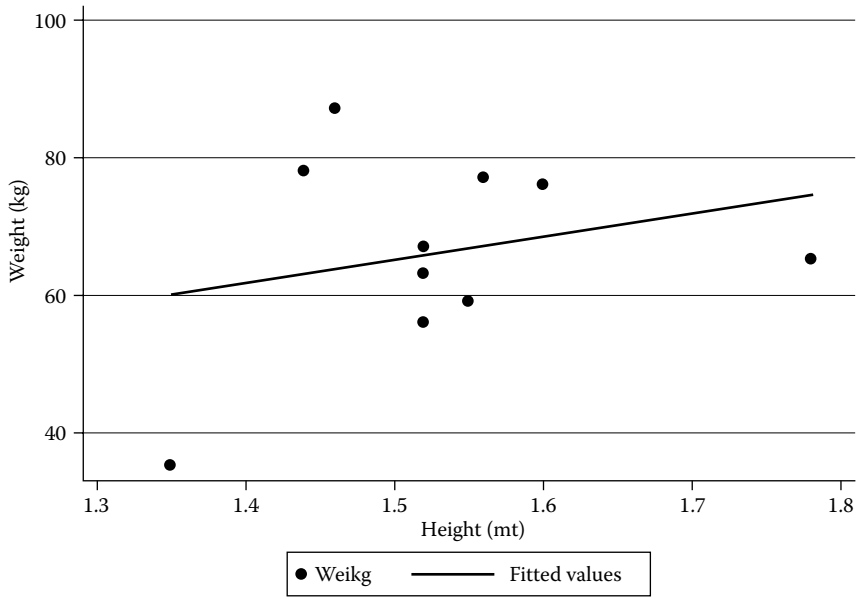


**Figure 5.1 Scatter plot.**

## 5.8 Running the Model

The results of this scatter plot show a linear upward trend between height and weight. Specifically, the higher the subjects, the heavier their weight. After generating a scatter plot, the user can run a linear regression model with these data using the command *regress* (or *reg*), as follows:

```
reg weikg heimt
```

*Output*

```
  Source |       SS       df       MS              Number of obs =      10
---------+------------------------------           F(1, 8)       =    0.56
   Model |  125.560422     1  125.560422           Prob > F      =  0.4765
Residual |  1800.53958     8  225.067447           R-squared     =  0.0652
---------+------------------------------           Adj R-squared = -0.0517
   Total |     1926.1     9  214.011111           Root MSE      =  15.002

------------------------------------------------------------------------------
   weikg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
   heimt |   33.12939   44.35508     0.75   0.476    -69.15362    135.4124
   _cons |   15.61203     68.0289     0.23   0.824    -141.2629     172.487
------------------------------------------------------------------------------
```

The results of this command show two tables. The first table describes the estimated ratio of the MS obtained by the model over the residual MS is less than 1 $\left[(125.6/225.1)=0.56\right]$. This result indicates that there is no evidence to reject the null hypothesis (*P*-value = .4765 > .05). The second table shows the estimated coefficients of the model for the predictor weight and for the intercept (_cons); so, the linear trend is estimated using the following equation: $\widehat{\text{Weight}} = 15.6 + 33.1 * \text{height}$. Thus, the estimated expected weight in kilograms will increase 33.1 (95% CI: –69.2, 135.4) for every additional meter of height. However, this increasing trend was not significant (*P*-value > .05). The percentage of total variation from $\bar{Y}$ explained by the model is 6.5% (*R*-squared). In this case, Student's *t*-test described below the ANOVA table shows nonsignificant results for the predictor *weight* with exactly the same *P*-value described for the *F*-distribution in ANOVA, which is because of the fact that in an SLRM, $t^2 = F$.

## 5.9 Centering

To facilitate the interpretation of the intercept on a linear regression model, it is advisable to transform the values of $X_i$ to the difference of each value from its mean as $\left(X_i - \bar{X}_i\right)$. This transformation is known as *centering*. As a result of the

centralization, the estimator of the coefficient associated with the intercept is equal to the mean of the dependent variable, that is,

$$\hat{\beta}_0 = \bar{Y}$$

This process does not affect the estimates of the coefficients associated with the independent variable. Assuming the previous database, the process of centering height to explain weight in STATA can be achieved by typing the following commands:

```
sum weikg
sum heimt
*Centering weikg using the result of the previous sum command
gen heimtc=heimt-r(mean)
reg weikg heimtc
```

*Output*

```
sum weikg
  Variable |     Obs    Mean     Std. Dev.   Min        Max
-----------+--------------------------------------------------
     weikg |      10    66.3      14.62912    35         87

.          sum heimt
  Variable |     Obs    Mean     Std. Dev.   Min        Max
-----------+--------------------------------------------------
     heimt |      10    1.53      .1127435   1.35       1.78

*Centering weikg using the result of the previous sum command

gen heimtc=heimt-r(mean)

reg weikg heimtc
    Source |    SS     df      MS      Number of obs =      10
-----------+------------------------   F(1, 8)       =    0.56
     Model | 125.56044  1  125.56044   Prob > F      =  0.4765
  Residual |1800.53956  8 225.067445   R-squared     =  0.0652
-----------+---------------            Adj R-squared = -0.0517
     Total |    1926.1  9 214.011111   Root MSE      =  15.002
------------------------------------------------------------------------------
     weikg |   Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
    heimtc | 33.12939  44.35508     0.75   0.476    -69.15361   135.4124
     _cons |     66.3  4.744127    13.98   0.000     55.36002   77.23998
------------------------------------------------------------------------------
```

Now the equation model can be expressed in the following manner:

$$\widehat{\text{weikg}} = 66.3 + 33.1 * \text{heimtc} = 66.3 + 33.1 * (\text{heimt} - 1.53)$$

where $\widehat{\text{weikg}}$ is the estimated expected weight (kilograms), explained by centering height (meters). The user can confirm that the output is the same with the exception of the estimated constant coefficient ($\hat{\beta}_0 = 66.3$), which is the mean of the variable *weight* and is identified in Stata in the row labeled *_cons.*



## 5.10 Bootstrapping

Bootstrapping is a robust alternative to classical statistical methods when the assumptions are not met using these methods; it provides more accurate inferences, particularly when the sample size is small. The procedure to perform bootstrapping is via resampling methods for estimating standard errors and computing confidence intervals (Good, 2006). Bootstrapping in Stata can be done using the option *vce(boot)* in the command *reg*. For example, assuming the previous database, the command for estimating weight with centering height using bootstrapping estimates is as follows:

```
reg weikg heimtc, vce(boot)
```

*Output*

```
Bootstrap replications (50)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50

Linear regression                  Number of obs     =         10
                                   Replications      =         50
                                   Wald chi2(1)      =       0.24
                                   Prob > chi2       =     0.6275
                                   R-squared         =     0.0652
                                   Adj R-squared     =    -0.0517
                                   Root MSE          =    15.0022
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
       weikg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      heimtc |   33.12939    68.2667     0.49   0.627    -100.6709    166.9297
       _cons |        66.3   4.538575    14.61   0.000     57.40456    75.19544
------------------------------------------------------------------------------
```

The results show that the estimates of the regression coefficients are the same as those obtained using the least-squares method, but the standard errors for the coefficient of heimtc are different, being 44.4 versus 68.3. It is likely that these differences are due to the small sample size that was used in this example. For more information on this topic, we recommend checking out the book by Draper and Smith (1998).

## 5.11 Multiple Linear Regression Model

An extension of the SLRM is to use more predictor variables to improve the estimation of the expected value of *Y*. This model extension is called a multiple linear regression model (MLRM), and it is represented as follows for *m*-predictors:

$$\mu_{y/x} = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$$

where:

$\mu_{y/x}$ indicates the expected value of *Y* explained by the *X* variables for the *i*th subject
$X_j$ indicates the predictor variables ($j = 1,\ldots, m$)
$\beta_j$ indicates the coefficient (constant) associated with $X_j$

The assessment of the overall significance of its regression coefficients ($\beta_i$s, $i > 0$) can be performed using an ANOVA table, as follows:

| *Source of Variation* | *SS* | *df* | *MS* | F-*Ratio* |
|---|---|---|---|---|
| Regression | SSR | *m* | MSR = SSR/*m* | $F_c$ = MSR/MSE |
| Residual | SSE | $n - m - 1$ | MSE = SSE/($n - m - 1$) | |
| Total | TSS | $n - 1$ | MST = TSS/($n - 1$) | |

$$\text{TSS} = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$$

$$\text{SSR} = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$$

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

where:

$\hat{y}_i$ indicates the estimated expected value of *Y* given a set of specific values of the predictors
*X*s for the *i*th subject, as follows: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_m X_m$
$\hat{\beta}_j$ indicates the estimated value of the coefficient $\beta_j$
$\bar{y}_i$ indicates the overall mean of *Y*

For a multiple regression model, the ANOVA table assumes that $H_0$: $\beta_1 = \beta_2 = \ldots = \beta_m = 0$. If the calculated value of the test statistic $F_c$ is greater than $F_{(1-\alpha;\, m,\, n-m-1)}$ for a given significance level $\alpha$, we conclude that there is evidence against $H_0$.

In MLRMs, the Stata command *reg* can be used in a manner that is similar to how the simple linear regression was programmed, except that in the latter, the model consists of more than one independent variable. For example, assuming the previous database, to explain the expected *bmi* by the predictors *age* and *sex,* the specifications of the *reg* command are as follows:

```
reg  bmi age sex
```

*Output*

```
   Source |       SS       df       MS              Number of obs =      10
----------+------------------------------           F(  2,     7) =    2.81
    Model |  192.330078     2  96.1650389            Prob > F      =  0.1269
 Residual |   239.30065     7  34.1858072            R-squared     =  0.4456
----------+------------------------------           Adj R-squared =  0.2872
    Total |  431.630728     9  47.9589698            Root MSE      =  5.8469

------------------------------------------------------------------------------
       bmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
       age |  -.4433135   .3941954    -1.12   0.298    -1.375438    .4888106
       sex |   10.47109   4.432552     2.36   0.050    -.0102331    20.95241
     _cons |   36.82826    11.1896     3.29   0.013     10.36907    63.28745
------------------------------------------------------------------------------
```

The results show that the fitted MLRM is

$$\widehat{\text{bmi}} = 36.83 - 0.44 * \text{age} + 10.47 * \text{sex}$$

with *adjusted* $\hat{R}^2 = 28.72\%$. The overall assessment of ANOVA indicates that we are unable to reject the null hypothesis ($H_0: \beta_{\text{age}} = \beta_{\text{sex}} = 0$); thus, this result suggests that the coefficients of the predictors are equal to zero ($P$-value > .05), but the $t$-test for the predictor *sex* suggests that its coefficient in the model could be different from zero ($P$-value = .05). This contradictory result could be explained by the linear dependency (or *multicollinearity*) between the predictors, which affects the procedure that is used to estimate the coefficients of the MLRM. For example, if we want to run the SLRM of each predictor, bmi explained by age and bmi explained by sex, the resulting equations are as follows:

$$\widehat{\text{bmi}} = 26.30 + 0.07 * \text{age}$$

and

$$\widehat{\text{bmi}} = 24.59 + 7.72 * \text{sex}$$

When we compare the coefficient estimates of these equations with the equation of the MLRM, we can see different estimates in the coefficient values: 0.07 vs. −0.44 in *age* and 7.72 vs. 10.47 in *sex*. The presence of one of them affects the estimate of the coefficient associated with the other predictor. Unless both predictors are completely independent of each other, the coefficient estimates will not be affected by the presence or absence of one of them (Draper and Smith, 1998).



## 5.12 Partial Hypothesis

When ANOVA results are significant in MLRMs, the user may wish to determine which set of specific predictor variables are the most significant when all the predictors are assessed simultaneously. For this evaluation, it is recommended that the additional sum of squares of the residuals in the model without these predictors (*incomplete model*) be compared, using the model with these predictors (*complete model*) as a reference. When we reduce the predictor variables from a linear regression model, the sum of squares of the residuals increases. For example, the residual sum of squares for bmi explained by different predictors is described in the following table:

| *Predictors in the Model* | *Residual Degrees of Freedom* | *Residual Sum of Squares* | *Additional Sum of Squares* |
|---|---|---|---|
| Age + sex **(complete model)** | 7 | 239.3 | – |
| Age **(incomplete model)** | 8 | 430.1 | 190.8 |
| Sex **(incomplete model)** | 8 | 282.5 | 43.2 |

In both incomplete models, the additional sum of squares increases. To assess if this increment is statistically significant, a partial *F*-statistic is used. For example, let us assume the following notation for the complete and incomplete models:

*Complete Model*: $\mu_{y|X} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \beta_{k+1} X_{k+1} + \ldots + \beta_m X_m$

*Nested or Incomplete Model*: $\mu_{y|X'} = \beta'_0 + \beta'_1 X_1 + \ldots + \beta'_k X_k$

The user has to be aware that the coefficients from both models do not necessarily have the same value. Based on these models, a partial hypothesis can be defined with the following equation:

$$H_0 : \beta_{k+1} = \beta_{k+2} = \ldots = \beta_m = 0 \big|_{X_1, X_2, \ldots, X_k}$$

Then, the following steps are performed to evaluate this type of partial hypothesis:

1. Calculate the sum of squares of the residuals in the complete model ($SSE_{com}$) with $n - m - 1$ degrees of freedom.
2. Calculate the sum of squares of the residuals in the incomplete model ($SSE_{inc}$) with $n - k - 1$ degrees of freedom.
3. Compute the difference of SS between the sum of squares of the complete and incomplete models, which is called the additional sum of squares, with $m - k$ degrees of freedom.
4. Compute the following formula (partial *F*):

$$F\left(X_{k+1}, \ldots, X_m | X_1, \ldots, X_k\right) = \frac{\left(\mathrm{SSE}_{\mathrm{inc}} - \mathrm{SSE}_{\mathrm{com}}\right) / \left(m - k\right)}{\mathrm{SSE}_{\mathrm{com}} / \left(n - m - 1\right)}$$

5. Calculate the *P*-value using Fisher's *F*-distribution with $m - k$ and $n - m - 1$ degrees of freedom.

Considering the previous data of the sum of squares, the partial *F,* discarding *sex* from the complete model, will be:

$$F\left(\mathrm{sex} \mid \mathrm{age}\right) = \frac{190.8/1}{239.3/7} = 5.58$$

And the partial *F,* again discarding *age* from the complete model, will be:

$$F\left(\mathrm{age} \mid \mathrm{sex}\right) = \frac{43.2/1}{239.3/7} = 1.26$$

The respective null hypotheses are

$$H_0 : \beta_{\mathrm{sex}|\mathrm{age}} = 0$$

and

$$H_0 : \beta_{\mathrm{age}|\mathrm{sex}} = 0$$

The respective *P*-values are computed with the *F*-Fisher probability distribution with 1 and 7 degrees of freedom. In Stata, these *P*-values can be obtained using the *Ftail* command, as is illustrated in the following:

```
. dis Ftail(1,7,(282.536554 - 239.30065)/34.1858072)
0.29783641
. dis Ftail(1,7,(430.07574- 239.30065)/34.1858072)
0.05017009
```

An alternative procedure is to use the *test* command after the *reg* command for the complete model, as in the following:

```
quietly: reg bmi age sex
test sex
test age
```

*Output*

```
. test sex

( 1)  sex = 0

      F(  1,     7) =    5.58
           Prob > F =    0.0502

. test age

( 1)  age = 0

      F(  1,     7) =    1.26
           Prob > F =    0.2978
```

Thus, we conclude that there is marginal evidence against the null hypothesis, $H_0$: $\beta_{sex|age} = 0$ (*P*-value = .05), suggesting that the variable *sex* could be part of the model when the variable *age* is already one of the predictors.

## 5.13 Prediction

Should the user pursue using the model for predicting the expected value under the specific conditions of the predictors, the *adjust* command is available in Stata for this purpose. For example, assuming that the user is interested in estimating the expected bmi for females (sex = 1) aged 30 years; after the *reg* command, the specifications for the adjust command are as follows:

```
quietly: reg bmi age sex
adjust age=30 sex=1, ci
```

*Output*

```
------------------------------------------------------------------------------
    Dependent variable: bmi     Command: regress
Covariates set to value: age = 30, sex = 1
------------------------------------------------------------------------------

-----------------------------------------------
     All |         xb          lb           ub
---------+-------------------------------------
         |    33.9999    [26.8742     41.1257]
-----------------------------------------------
    Key:  xb         =  Linear Prediction
          [lb , ub]  =  [95% Confidence Interval]
```

The option *ci* in the adjust command is used to display the 95% confidence interval of the prediction. The results displayed in the above table indicate that for 30-year-old females, the estimated expected bmi is 34 (95% CI: 26.9, 41.1).

## 5.14 Polynomial Linear Regression Model

Another extension of the SLRM is the polynomial linear regression model, with at least one predictor at the power greater than 1. This model is recommended when it is suspected that a nonlinear trend would better explain the relationship between the outcome of interest and the predictors. The simplest polynomial model is the following expression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + e_i$$

The model above is known as a second-order or quadratic polynomial model, because it contains an independent variable expressed as a term to the first power ($X_i$) and a term expressed to the second power ($X_i^2$). To use this model, it is recommended that all predictors be centralized to reduce the effect of the correlation among predictors. For example, to explain the variable *bmi* by *age* and *age*$^2$, the following commands generate the estimates of the expected values for both the linear and polynomial models:

```
quietly: sum age
gen agec=age-r(mean)
gen agec2=agec^2
reg bmi agec
predict bmiesp1
reg bmi agec agec2
predict bmiesp2
```

*Output*

```
quietly: sum age
. gen agec=age-r(mean)
. gen agec2=agec^2
. reg bmi agec
```

| Source | SS | df | MS | Number of obs = | 10 |
|---|---|---|---|---|---|
| Model | 1.55497935 | 1 | 1.55497935 | F( 1, 8) = | 0.03 |
| Residual | 430.075749 | 8 | 53.7594686 | Prob > F = | 0.8692 |
| | | | | *R*-squared = | 0.0036 |
| Total | 431.630728 | 9 | 47.9589698 | Adj *R*-squared = | -0.1209 |
| | | | | Root MSE = | 7.3321 |

| bmi | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| agec | .0701375 | .4123967 | 0.17 | 0.869 | -.8808511 | 1.021126 |
| _cons | 28.45408 | 2.318609 | 12.27 | 0.000 | 23.10736 | 33.8008 |

```
. predict bmiesp1
(option xb assumed; fitted values)

. reg bmi agec agec2
```

| Source | SS | df | MS | Number of obs = | 10 |
|---|---|---|---|---|---|
| Model | 68.2219892 | 2 | 34.1109946 | F( 2, 7) = | 0.66 |
| Residual | 363.408739 | 7 | 51.9155342 | Prob > F = | 0.5476 |
| | | | | *R*-squared = | 0.1581 |
| Total | 431.630728 | 9 | 47.9589698 | Adj *R*-squared = | -0.0825 |
| | | | | Root MSE = | 7.2052 |

| bmi | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| agec | .7288876 | .7086385 | 1.03 | 0.338 | -.9467761 | 2.404551 |
| agec2 | -.0769584 | .0679124 | -1.13 | 0.294 | -.2375457 | .0836289 |
| _cons | 30.88673 | 3.130483 | 9.87 | 0.000 | 23.48432 | 38.28915 |

```
. predict bmiesp2
(option xb assumed; fitted values)
```

Once the expected value for each model is estimated with the command predict (bmiesp1 and bmiesp2), a plot with these estimates can be displayed with the following command (Figure 5.2):

```
twoway (scatter bmi age, sort) (line bmiesp1 age, sort) (line
bmiesp2 age, sort), ytitle(bmi)
```

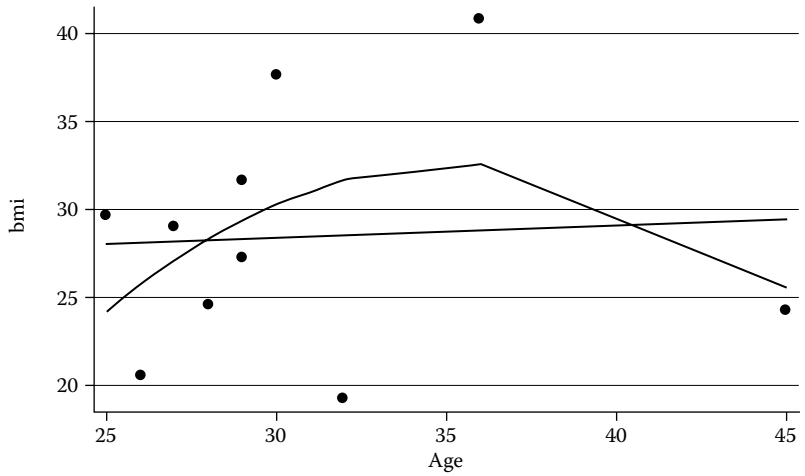


**Figure 5.2 Linear and polynomial fits.**

*Output*

In this example, the quadratic curve appears to be better than the linear trend in terms of its ability to explain the expected value of bmi by *age*.

## 5.15 Sample Size and Statistical Power

To determine the minimum sample size for performing an SLRM with enough statistical power (i.e., $1 - \beta = 0.8$) and a minimum significance level (i.e., $\alpha = 0.05$), the following expression is used (Kleinbaum et al., 2008):

$$n_s \geq \left[ \frac{Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}}{C(\rho)} \right]^2 + 3$$

in which $C(\rho)$ is the Fisher transformation, defined as follows:

$$C(\rho) = \frac{1}{2} \mathrm{Ln}\left( \frac{1+\rho}{1-\rho} \right)$$

In the above, $\rho$ can be estimated with the square root of the expected coefficient of determination ($R^2$) for the model under consideration.

In Stata we can estimate sample size using the option of correlation in the *Power and sample size analysis* window in the *Statistics* menu, providing significance level, power value, and the linear correlation coefficient value under the alternative

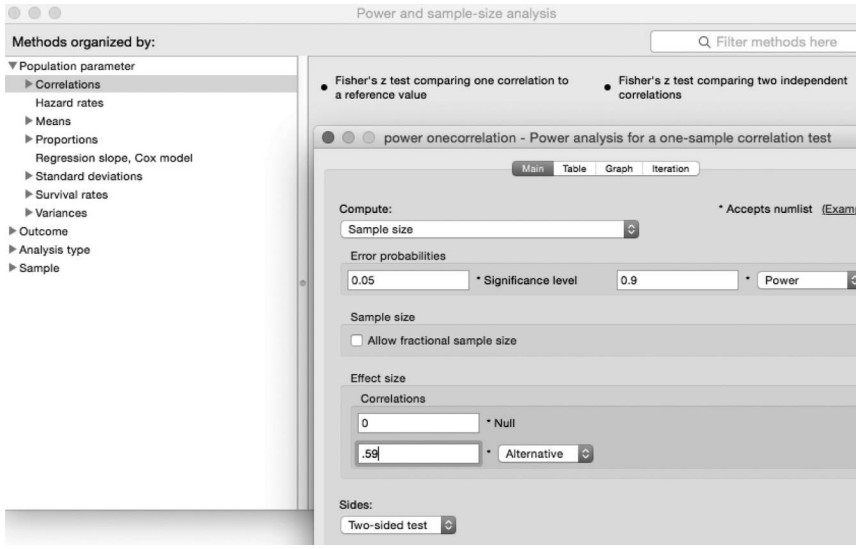


**Figure 5.3 Power and sample-size analysis.**

hypothesis. For example, assuming that we want to determine the minimum sample size needed to estimate the expected *bmi* value using an SLRM with sex as predictor and an approximate $R^2$ of 0.3454 ($\rho \sim \sqrt{.3454} = .59$), the dialog box should be filled out as described in Figure 5.3.

*Output*

```
Estimated sample size for a one-sample correlation test
Fisher's z test
Ho: r = r0  versus  Ha: r != r0

Study parameters:

        alpha =    0.0500
        power =    0.9000
        delta =    0.5900
           r0 =    0.0000
           ra =    0.5900
Estimated sample size:

            N =        26
```

Therefore, the minimum sample size for performing an SLRM between BMI and *sex,* assuming 90% statistical power and a 5% significance level, is 26 subjects. Should the user want to determine the minimum sample size for an MLRM ($n_m$), when $X_1$ is the main predictor, the following expression is recommended (Kleinbaum et al., 2008):

$$n_m = \frac{n_s}{1 - \rho^2_{X_1(X_2,\ldots,X_k)}}$$

where $n_s$ is the minimum sample size for the SLRM using $X_1$ as predictor, $\rho^2_{X_1(X_2,\ldots,X_k)}$ is the chosen value of the population-squared multiple correlation between the main predictor $X_1$ and the control variables $X_2$, $X_3$,…,$X_k$.



## 5.16 Considerations for the Assumptions of the Linear Regression Model

Having defined the linear regression model most suitable to your needs, it is necessary to verify its compliance with the assumptions for its creation. This assessment is conducted primarily through the absolute difference between the observed and expected values under the model:

$$\hat{e}_i = y_i - \hat{y}_i$$

where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

At first it is assumed that the residuals are independent; however, the $e_i$ obtained from the study data depend on the expected values of $Y$ under the model, which in turn depend on the values of the predictors. Moreover, the model assumes that

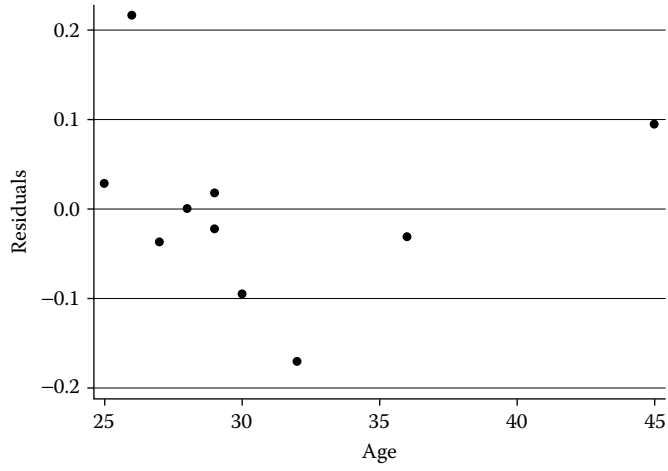


**Figure 5.4 Residuals distribution.**

the variances are constant, but the variance of the residual depends on the distance of the central values of the predictors. To verify the compliance of constant variance, it is recommended that the standardized residuals be graphically represented. In Stata, this type of graph can be obtained using the *rvpplot* command after the *reg* command. For example, to describe the residuals distribution related with the linear regression model between *heimt* and *age,* the Stata commands are:

```
reg heimt age
rvpplot age
```

The output of these commands can be seen in Figure 5.4. Because of the small sample size in this example, it is difficult to visualize a particular pattern around 0, although some symmetric distribution is observed. For more discussion on this topic, we recommend checking out the book by Draper and Smith (1998).