Polynomial terms in linear regression models add considerable flexibility to the estimated relationship between an explanatory variable and the dependent variable. For example in a cross section model of wages, the relationship between the ages of a survey respondent and the wages they report is very often not linear. Typically, hourly wages rise quite quickly at young ages as labor market experience is rewarded with wage increases. After a certain age, often around 50 years of age, reported wages tend to decline as age increases. At the peak, one might say that the combination of experience and energy maximize hourly wages. But after that, for a number of reasons, reported wages decline as age increases (perhaps very high wage earners retire earlier, perhaps energy levels decline so people who work on commissions earn less per hour compared to their peak years.)

Choosing the Appropriate Polynomial

The following model has been estimated by least squares using a sample drawn from Canada’s Labour Force Survey. A, A2 and A3 are the age, square of age and the cube of age. Years of formal education is recorded by E, its square E2 and cube E3. The dummy variable f equals 1 for females and 0 for males.

Table 1: Cubic Polynomials

Call:
\texttt{lm(formula = \textit{wage} \sim A + A2 + A3 + E + E2 + E3 + f)}

Coefficients:

|        | Estimate | Std. Error | t value | Pr(>|t|) |
|--------|----------|------------|---------|----------|
| (Intercept) | -6.104e+00 | 1.847e+01  | -0.331  | 0.741034 |
| A      | 1.079e+00  | 2.871e-01  | 3.760   | 0.000172 *** |
| A2     | -1.282e-02 | 7.344e-03  | -1.745  | 0.080995 . |
| A3     | 2.950e-05  | 5.921e-05  | 0.498   | 0.618320  |
| E      | -8.256e-02 | 4.209e+00  | -0.020  | 0.984353  |
| E2     | -2.341e-02 | 3.236e-01  | -0.072  | 0.942332  |
| E3     | 4.569e-03  | 8.115e-03  | 0.563   | 0.573386  |
| f      | -4.494e+00 | 3.032e-01  | -14.819 | < 2e-16 *** |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.477 on 4022 degrees of freedom
Multiple R-squared: 0.2677,    Adjusted R-squared: 0.2664
Note that the female dummy variable has a very large t-statistic and so the hypothesis that the
ture coefficient is zero is clearly rejected even at the 1% level of significance. It seems women
with the same education as men and at any given (common) age earn $4.49 less per hour than
males – although this model does not control for industry or job classification.

However, of the six coefficients on the age and education variables only one is statistically
significantly different from zero – the linear term on age. Does this mean that the data imply
wages are a linear function of age and wages do not depend on years of education at all? The
short answer is “no.” A more appropriate interpretation is that the relationship between wages
and age and education may well be non-linear, but the data do not require a shape that is as
flexible as a cubic function.

The usual strategy when one sees results like those in Table 1 is to reduce the order of the
polynomials to quadratics by deleting the cubic terms. The results are shown in Table 2.

Table 2: Quadratic Polynomials

Call:
  lm(formula = wage ~ A + A2 + E + E2 + f)

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.2637656  4.1538789  1.267    0.205
A            0.9387520  0.0702901 13.355  < 2e-16 ***
A2          -0.0091655  0.0008766 -10.455  < 2e-16 ***
E           -2.4225717  0.6116939  -3.960 7.61e-05 ***
E2           0.1583829  0.0228884   6.920 5.24e-12 ***
f          -4.4999268  0.3028190  -14.860  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.475 on 4024 degrees of freedom
Multiple R-squared:  0.2676, Adjusted R-squared:  0.2667

The elimination of the two cubic terms A3 and E3 hardly changes the R-squared at all – it falls
by 0.0001. We might have expected this given the small t-statistics on the coefficients of the
eliminated variables.

More interesting is the fact that the t-statistics on all the coefficients in Table 2 are now larger in
absolute value than the critical value for a test at the 1% level of significance. That critical value
is provided by R:

> qt(0.995,4024)
The implications of the results in Table 1 and Table 2 are that this sample suggests that the relationship between wages and age is adequately represented by a quadratic equation but that the more flexible cubic equation is not necessary. However, the t-statistics in Table 2 imply that a linear relationship between wages and age can be ruled out (the t-statistic on A2’s coefficient is 10.46) so a quadratic shape is required. The same can be said about the relationship between wages and years of formal education.

A remaining question is why, in Table 1, were the linear terms E and E2 not statistically significant just because E3 was included? The answer is that E, E2 and E3 are all highly correlated so deleting any one of them and leaving the other two will provide a good fit to the data. The individual t-statistics being so close to zero imply that any one of the three variables could be deleted with little effect on R-squared. But these t-statistics do not indicate what would happen if two or all three were deleted together. Note the correlations are:

```r
> cor(E,E2)
 [1] 0.9932369
> cor(E,E3)
 [1] 0.9748923
> cor(E2,E3)
 [1] 0.9940881
```

However, when one of the three is deleted, the results (Table 2) show that the other two must be retained. Another question is: why, given the t-statistics in Table 1, did we not choose to drop E since its coefficient had the smallest t-statistic in absolute value? If we had dropped E and retained E2 and E3 we would have retained a cubic equation with no linear term. This is certainly an option, but usually researchers opt for the lowest order polynomial that accounts for the data. On these grounds the leading term E3 was dropped.

**Interpretation of the Quadratic Relationships**

In Table 2 the relationship between age and wages is quadratic. In general, a quadratic can be represented as follows:

\[ y = a + bx + cx^2 \quad [1] \]

Quadratic equations are always symmetric about an axis parallel to the y-axis. The axis of symmetry passes through the maximum or minimum point. If \( c > 0 \), there is a minimum point and if \( c < 0 \) there is a maximum point. The extreme point is located by setting the first derivative to zero:
\[
dy{dx} = b + 2cx = 0 \Rightarrow x^* = \frac{-b}{2c} \quad [2]
\]

Note that the extreme point is at a positive value of \(x\) if \(b\) and \(c\) have opposite signs as they do in both cases shown in Table 2.

Equation [1] can be rewritten to reveal its symmetry about the vertical axis at \(x^*\):

\[
y = c \left( \frac{a}{c} + \frac{b}{c}x + x^2 \right) = c \left( \frac{a}{c} + \left( x - \frac{b}{2c} \right)^2 - \frac{b^2}{4c^2} \right) = \left( a - \frac{b^2}{4c} \right) + \left( x - x^* \right)^2 \quad [3]
\]

Equation [3] clearly shows that \(y\) is symmetric about \(x^*\) and that at the extreme point, the value of \(y\) is:

\[
y^* = \left( a - \frac{b^2}{4c} \right) \quad [4]
\]

Equation [4] can be confirmed by substituting \(x^*\) from [2] into [1] although it is immediately obvious from [3].

**Wages and Age**

The relationship between predicted wages and age that is implied by the results in Table 2 can be written as:

\[
\hat{w} = a + 0.94age - 0.0092age^2
\]

The first derivative shows how predicted wages change as age increases:

\[
d\hat{w}/d(age) = 0.94 - 0.0184age \quad [5]
\]

Since the coefficient on age squared is negative, the function has a maximum value and this maximum value occurs at (see [2]):

\[
\hat{w}^* = \frac{0.94}{2 \times 0.0092} = 51.09 \approx 51 \text{ years}
\]

Up to 51 years of age, predicted wages increase as age increases but at a diminishing rate (see equation [5]). After 51 years of age, wages decline at an increasing rate.
For example, at age = 20, predicted wages rise at the rate of (0.94 - 0.0184x20) = $0.572 per hour per year i.e. 57.2 cents per hour as age increases by one year. At age = 40, hourly wages are predicted to rise by just (0.94 - 0.0184x40) = $0.204 per hour per year.

At age = 51 wages are not rising at all and start to fall as age goes beyond 51 years.

Perhaps the easiest way to show the relationship between wages and age is to create a plot over the age range 20 to 65 years. This can be done quite easily in R. First create a vector with values that span the required age range. Give this vector a unique name, say AA, and certainly do not use the same name as the label that defines the age data. The predicted wages must be placed in another vector that has a previously unused name, say WW.

```
> AA <- seq(20,65,1)
> WW <- 0.94*AA-0.0092*AA^2
> plot(AA,WW,type="l",lwd=3)
```

![Graph](image.png)

The shape of the function shown above is an accurate representation of the quadratic equation (note the peak at 51 years of age). However, the vertical axis is not a sensible predicted wage since no attempt was made to identify an appropriate intercept. One way to do this is to ensure the plot predicts the mean wage when the mean age is substituted into the equation:

$$\bar{w} = a + 0.94(\bar{age}) - 0.0092(\bar{age}^2)$$

Use R to determine the appropriate intercept as follows:
> intercept <- mean(wage) - 0.94*mean(age)+0.0092*(mean(age)^2)
> intercept
[1] -1.639968
> ones <- seq(1,1,length=46)
> WW <- intercept*ones+0.94*AA-0.0092*AA^2
> plot(AA,WW,xlab="Age",ylab="Predicted wage",type="l",lwd=3,main="Plot of Wages and Age")

In the second plot, the predicted wage at the mean age is clearly between $20 and $22 per hour. By calculation, R shows the mean wage in this sample is $21.05 per hour.

A similar method can be used to create a plot of wage versus years of formal education based on the results in Table 2. A chart appears below along with the R-code:

> int2 <- mean(wage)+2.42*mean(educ)-0.158*(mean(educ)^2)
> int2
[1] 25.21263
> summary(educ)
 Min. 1st Qu.    Median  Mean 3rd Qu.   Max. 
   8.00  12.00  14.00   13.34  14.00  18.00
> EE <- seq(8,18,1)
> ones2 <- seq(1,1,length=11)
> WWW <- int2*ones2 - 2.42*EE + 0.158*EE^2
> plot(EE, WWW, xlab="Years of Education", ylab="Predicted Wages", main="Wages and Education", type="l", lwd=3)