12. POLYNOMIAL REGRESSION

The cure for boredom is curiosity. There is no cure for curiosity.

– Ellen Parr

Polynomial regression models are attractive for fitting data because their shape is so malleable. By adding higher-order terms and changing the signs and magnitudes of the coefficients, a variety of complex curve shapes can be obtained. A polynomial regression model has the form

\[ y = \hat{a}_0 + \hat{a}_1 x + \hat{a}_2 x^2 + \hat{a}_3 x^3 + \ldots + \hat{a}_k x^k + e \]

If we make the following substitutions:

\[ x_1 = x \quad , \quad x_2 = x^2 \quad , \quad x_3 = x^3 \quad , \quad \ldots \quad , \quad x_k = x^k \]

we get a model,

\[ y = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \ldots + \hat{a}_k x_k + e \]

that is identical to the multiple regression model we studied in the last section. The difference is simply that the \( k \) independent regressor variables have been replaced by \( k \) different powers of a single variable \( x \).

Note that our polynomial regression model is still a linear regression model because it is a linear function of the regression coefficients. So the regression coefficients can be found using the same least-squares approach we’ve been using. We’ll once again assume that the model residual \( e \) is a normally-distributed random error with a mean of zero and an unknown variance \( \sigma^2 \) that is constant for all values of \( x \). To minimize \( e \) for the entire sample set, we derive an equation for the sum of the squares of the residuals,

\[ S_r = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} \left( y_i - \hat{a}_0 - \hat{a}_1 x_i - \hat{a}_2 x_i^2 - \hat{a}_3 x_i^3 - \ldots - \hat{a}_k x_i^k \right)^2 \]

then set the partial derivatives of that equation to zero,

\[ \frac{\partial S_r}{\partial a_0} = 0 \quad , \quad \frac{\partial S_r}{\partial a_1} = 0 \quad , \quad \frac{\partial S_r}{\partial a_2} = 0 \quad , \quad \frac{\partial S_r}{\partial a_3} = 0 \quad , \quad \ldots \quad , \quad \frac{\partial S_r}{\partial a_k} = 0 \]

and solve the resulting system of simultaneous linear equations (that is, linear functions of the regression coefficients) for \( a_0, a_1, a_2, a_3, \ldots, a_n \).

In this course, we’ll only deal with polynomial functions of a single regressor variable \( x \), but it is also possible to fit a polynomial function with two or more independent regressor variables:

\[ y = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \hat{a}_3 x_1^2 + \hat{a}_4 x_2^2 + \ldots + \hat{a}_k x_1^k + \hat{a}_l x_2^k + e \]

as long as the model is a linear function of the regression coefficients.
**Example**

When cement is mixed with water, the calcium silicates in the cement combine with the water to create calcium silicate hydrate (CSH) gel and calcium hydroxide (CH) crystals. The CSH is what gives the hardened cement paste its strength. The CH is a byproduct that just takes up space. If we mix fly ash (a waste product from coal-fired power plants) in with the cement, it reacts with the CH to produce more CSH. So, by blending inexpensive fly ash with the relatively expensive cement we can get the same or higher strengths at less cost!

To determine how concrete strength varies with blending ratio, researchers made six batches of concrete with increasing percentages of fly ash. From each batch of concrete they made three test cylinders and cured them in water for 28 days. The 28-day cylinder strengths are tabulated below:

<table>
<thead>
<tr>
<th>Fly Ash (%)</th>
<th>Strength (psi)</th>
<th>Fly Ash (%)</th>
<th>Strength (psi)</th>
<th>Fly Ash (%)</th>
<th>Strength (psi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4779</td>
<td>20</td>
<td>5110</td>
<td>40</td>
<td>5746</td>
</tr>
<tr>
<td>0</td>
<td>4706</td>
<td>20</td>
<td>5685</td>
<td>40</td>
<td>5719</td>
</tr>
<tr>
<td>0</td>
<td>4350</td>
<td>20</td>
<td>5618</td>
<td>40</td>
<td>5782</td>
</tr>
<tr>
<td>10</td>
<td>5189</td>
<td>30</td>
<td>5995</td>
<td>50</td>
<td>4895</td>
</tr>
<tr>
<td>10</td>
<td>5140</td>
<td>30</td>
<td>5628</td>
<td>50</td>
<td>5030</td>
</tr>
<tr>
<td>10</td>
<td>4976</td>
<td>30</td>
<td>5897</td>
<td>50</td>
<td>4648</td>
</tr>
</tbody>
</table>

Plotting the data, we see a nonlinear relationship between strength and fly ash content:

The plot shows a definite peak at a fly ash content of about 30%, which suggests some type of quadratic model would be appropriate.
As before, we can enter this data into an Excel spreadsheet and use the LINEST function to solve for the regression coefficients (as well as many of the regression statistics):

|  -1.562 | 88.224 | 4498.190 |
| 0.229672815 | 11.9635803 | 127.1870458 |
| 0.786803152 | 243.0628608 | #N/A |
| 27.67875643 | 15 | #N/A |
| 3270497.186 | 886193.3143 | #N/A |

From this table, we observe that

\[ \hat{a}_0 = 4498.19, \quad \hat{a}_1 = 88.224, \quad \hat{a}_2 = -1.562 \]

so the quadratic regression model is:

\[ y = 4498.19 + 88.224x - 1.562x^2 + e \]

where \( y \) is the 28-day unconfined compression strength of the concrete and \( x \) is the percentage of fly ash blended with the cement to make the concrete.

We also observe that

\[ R^2 = 0.7868, \quad s_{y|x} = 243 \text{ psi} \]

so our model explains 78.7% of the variation in the measured strengths and the residuals have a standard deviation of 243 psi (which is a point estimate of the unknown \( \sigma \)).

The \( F \)-test statistic can be used to determine the significance of the regression:

\[ H_0: \quad a_1 = a_2 = 0 \]

\[ H_a: \quad a_1 \neq 0 \text{ and/or } a_2 \neq 0 \]

T.S.: \[ F = \frac{MSR}{MSE} = 27.68 \]

R.R.: \[ F > F_{a,k,n-k-1} = F_{0.05,2,15} = 3.68 \]
We can use the $t$-test to determine the contribution of $x$ and $x^2$ to the overall prediction, which will tell us whether or not the relationship is quadratic (i.e., nonlinear):

- $H_0$: $a_1 = 0$
- $H_a$: $a_1 \neq 0$

Test Statistic (T.S.): $t = \frac{\hat{a}_1}{s_1} = \frac{88.224}{11.964} = 7.37$

Reject the Null Hypothesis (R.R.): $|t| > t_{\alpha/2, n-k-1} = t_{0.025,15} = 2.13$

- $H_0$: $a_2 = 0$
- $H_a$: $a_2 \neq 0$

Test Statistic (T.S.): $t = \frac{\hat{a}_2}{s_2} = \frac{-1.562}{0.230} = -6.79$

Reject the Null Hypothesis (R.R.): $|t| > t_{\alpha/2, n-k-1} = t_{0.025,15} = 2.13$

The natural question to ask is whether a higher-order model might do a better job. This means adding additional terms such as $x^3$ and $x^4$ to the multiple regression model. One way to determine if more terms are better is to compare the $R^2$ values to see if the percentage of the data scatter explained by the model is increasing or decreasing. To make it a fair comparison, though, we have to use adjusted $R^2$ values:
A Few Caveats

Polynomial regression is a very powerful tool but it is very easy to misuse. In general, the order of the polynomial is one greater than the number of maxima or minima in the function. A straight line, for example, is a 1st-order polynomial and has no peaks or troughs. A parabola is a 2nd-order polynomial and has exactly one peak or trough. A cubic is a 3rd-order polynomial and has one peak and one trough.

If we add another term to our concrete strength model in order to improve the fit, there must be a trough somewhere since there’s a peak at approximately 30% fly ash. It might be to the left of the strength axis, in which case it won’t matter, but it could just as easily be at 55% fly ash, in which case the modeled concrete strength would go back up at higher fly ash contents.

For example, the best fit to the concrete strength data is obtained with a 5th-order polynomial model (which has an $R^2$ value of 0.885) but the behavior of the model beyond the maximum observed fly ash content of 50% is not at all what we would want or expect:

It is very important that you never extrapolate a polynomial regression model beyond the range of the observations. It’s a sure-fire way to get into trouble!