POLYNOMIAL AND MULTIPLE REGRESSION

Polynomial Regression

- Polynomial regression used to fit nonlinear (e.g. curvilinear) data into a least squares linear regression model.

- It is a form of linear regression that allows one to predict a single y variable by decomposing the x variable into a n\(^{th}\) order polynomial.

Generally has the form:

\[ y = a + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \ldots \beta_k x^k \]

- In polynomial regression, different powers of the x variable are successively added to the equation to see if they increase \(r^2\) significantly.

Note, as successive powers of x are added to the equation, the best fit line changes shape.

\[ y = a + \beta_1 x \] (a straight line)
\[ y = a + \beta_1 x + \beta_2 x^2 \] (this quadratic equation produces a parabola)
\[ y = a + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \] (this cubic equation produces a s-shaped curve)
An examination of a scatter plot of the data should help you decide to what power to use. For a single curvilinear line use the quadratic equation. If the curve has two bends, you'll need a cubic model. Curves with multiple kinks need even higher-order terms. It's rare to go past a quadratic.

**Significance tests** should be conducted to test if the polynomial regression explains a significant amount of the variance.

One conducts hypothesis tests at each step to see if the increase in $r^2$ is significantly different. Thus you test the difference in the $r^2$ value between the lower power and next higher power equations asking the question: does an additional term significantly increase the $r^2$ value?

Begin by testing if the data are linear. If so, stop (no polynomial regression is necessary). If the linear regression is insignificant, compare the $r^2$ for the linear versus the $r^2$ for the quadratic equation. Continue adding powers and doing tests until significance is reached.

$$H_0 : \quad r_i^2 = r_j^2$$

$$H_1 : \quad r_i^2 < r_j^2$$

Where $i$ refers to the lower power equation and $j$ refers to the higher power equation. This test is done using the F test statistic where:

$$F = df_j \frac{r_j^2 - r_i^2}{1 - r_j^2}$$

F is distributed with $j$ degrees of freedom in the numerator and $n-j-1$ degrees of freedom in the denominator where $j$ is the power of the higher order polynomial.

**Warning.** With the addition of successive powers of $x$, the $r^2$ will no doubt increase. As the powers of $x$ (k or the order of the polynomial equation) approaches $n-1$, the equation will begin to give a "perfect fit" by connecting all the data points. This is not what one wants!

Best to generally stay within the quadratic ($x^2$) or cubic ($x^3$) orders
Multiple Regression (Least Squares)

- Multiple regression allows one to predict a single variable from a set of $k$ additional variables where $k = m-1$

- Generally has the form:

$$y = a + \beta_1 x_1 + \beta_2 x_2 \ldots \beta_k x_k$$

  $a =$ initial intercept  
  $\beta =$ partial regression coefficients for variables $x_1$ .... $x_k$

- Can get the matrix of coefficients (the $\beta$'s) by the equation

$$B = A \cdot C$$

Where:

- $B =$ vector of regression coefficients
- $A =$ matrix of sums of squares and cross products (based on mean deviations)
- $C =$ a vector of sums of cross products of $y$

$$B = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} \Sigma x_i^2 & \ldots & \Sigma x_i x_k \\ \vdots & \ddots & \vdots \\ \Sigma x_k x_1 & \ldots & \Sigma x_k^2 \end{pmatrix} \begin{pmatrix} \Sigma x_i y \\ \vdots \\ \Sigma x_k y \end{pmatrix}$$

To get the initial intercept $(a)$ plug in the means for $y$ and each $x$, then solve for $a$

$$a = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 \ldots - \beta_k \bar{x}_k$$

The equation values for each regressed object onto the line equals:

$$\hat{y}_i = a + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots \beta_k x_{ik} + e_i$$
The error term \((e_i)\) for each \(\hat{y}_i\) can be found \((y_i - \hat{y}_i)\)

The \(\beta\)'s may be difficult to interpret (but may be thought of as weights or loadings).

In matrix form, \(Y = XB + E\) where ...

- \(Y\) is a vector of independent variables
- \(X\) is the matrix of dependent variables
- \(B\) is the column vector of regression coefficients
- \(E\) is the column vector of residuals (error terms)

- \(B\) is determined to minimize the sums of squares of the residuals and corresponds to finding the best fitting \(k\)-dimensional hyper plane in \(m\) dimensional space.

- The least-squares approach minimized the sum of the squared distances parallel to the \(y\)-axis

- Significant tests on multiple regression
  
  \[ H_0: \beta_0 = \beta_1 = \beta_2 \ldots \beta_k = 0 \]
  \[ H_1: \text{at least one of the } \beta \text{ coefficients } \neq 0 \]

- Statistics are the same as with bivariate regression and analysis of variance. Partition the variance into three different sums of squares

  \[ SS_r = \sum (\hat{y}_i - \bar{y})^2 \quad SS_D = \sum (y_i - \hat{y}_i)^2 \quad SS_T = SS_r + SS_D \]

- Variance explained by regression = \(S_r^2 = SS_r / k\)
- Unexplained variance = \(S_D^2 = SS_D / (n - k - 1)\)
- Total variance = \(S_t^2 = SS_t / (n - 1)\)

  note: the ratio \(SS_r/SS_t = R^2\) where \(R\) = the multiple correlation coefficient
• Use F-test to see if the regression explains a significant proportion of the variance, i.e. the multiple regression coefficients = 0. Observed value of F can be obtained by:

\[ F = \frac{S_r^2}{S_D^2} \]

and should be tested against a critical value of F from a table for a desired \( \alpha \) and \( k \) df in numerator and \( n-k-1 \) df in denominator.

• Each individual regression coefficient can also be tested by the t-test by calculating the standard error for each variable. For each coefficient calculate the standard error (se\(_b\)) and then calculated t value.

\[
\text{for } \beta \text{ coefficient } j = se_{\beta_j} = \sqrt{S_D^2 \cdot a_{jj}^{-1}}
\]

where \( a_{jj}^{-1} \) is element j in the inverses matrix of sums of squares and cross products

\[
t = \frac{\beta_j - 0}{se_{\beta_j}}
\]

test the observed t-value for each coefficient (which is distributed with \( n-k-1 \) degrees of freedom)
e.g. Variables measured on trilobites

<table>
<thead>
<tr>
<th>Sample</th>
<th>L (Y)</th>
<th>W(X₁)</th>
<th>GSL(X₂)</th>
<th>GA(X₃)</th>
<th>TH(X₄)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>6.2</td>
<td>4.05</td>
<td>1.3</td>
<td>5.05</td>
</tr>
<tr>
<td>2</td>
<td>24.2</td>
<td>6.45</td>
<td>4.3</td>
<td>1.25</td>
<td>5.25</td>
</tr>
<tr>
<td>3</td>
<td>24.9</td>
<td>6.95</td>
<td>3.9</td>
<td>1.15</td>
<td>5.7</td>
</tr>
<tr>
<td>4</td>
<td>24.65</td>
<td>6.75</td>
<td>4.1</td>
<td>1.2</td>
<td>5.65</td>
</tr>
<tr>
<td>5</td>
<td>24.75</td>
<td>7</td>
<td>4</td>
<td>1.2</td>
<td>5.3</td>
</tr>
<tr>
<td>6</td>
<td>25.5</td>
<td>7.15</td>
<td>4.1</td>
<td>1.2</td>
<td>5.6</td>
</tr>
<tr>
<td>7</td>
<td>25.3</td>
<td>7.35</td>
<td>4.05</td>
<td>1.3</td>
<td>5.85</td>
</tr>
<tr>
<td>8</td>
<td>24.75</td>
<td>6.85</td>
<td>4.1</td>
<td>1.2</td>
<td>5.55</td>
</tr>
<tr>
<td>9</td>
<td>25.05</td>
<td>6.85</td>
<td>4.35</td>
<td>1.35</td>
<td>5.5</td>
</tr>
<tr>
<td>10</td>
<td>25.15</td>
<td>6.85</td>
<td>4.1</td>
<td>1.25</td>
<td>5.65</td>
</tr>
</tbody>
</table>

**Model Summary**

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.974</td>
<td>.949</td>
<td>.907</td>
<td>.215</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), TH, CA, GSL, W

**ANOVAb**

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>4.275</td>
<td>4</td>
<td>1.069</td>
<td>23.062</td>
<td>.002a</td>
</tr>
<tr>
<td>Residual</td>
<td>.232</td>
<td>5</td>
<td>.046</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4.506</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), TH, CA, GSL, W

B. Dependent Variable: L

**Coefficientsb**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Sig.</th>
<th>95% Confidence Interval for β</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>1.767</td>
<td>.225</td>
<td>3.21</td>
<td>1.30</td>
</tr>
<tr>
<td>W</td>
<td>1.679</td>
<td>.377</td>
<td>.867</td>
<td>4.40</td>
</tr>
<tr>
<td>GSL</td>
<td>2.286</td>
<td>.709</td>
<td>.427</td>
<td>3.226</td>
</tr>
<tr>
<td>GA</td>
<td>-1.035</td>
<td>.502</td>
<td>-.264</td>
<td>-2.021</td>
</tr>
<tr>
<td>TH</td>
<td>.725</td>
<td>.107</td>
<td>.248</td>
<td>1.433</td>
</tr>
</tbody>
</table>

b. Dependent Variable: L