Polynomial Regression

POLYNOMIAL REGRESSION command fits a polynomial relationship between variables. The regression is estimated using ordinary least squares for a response variable and powers of a single predictor. Polynomial regression (also known as curvilinear regression) can be used as the simplest nonlinear approach to fit a non-linear relationship between variables. Polynomial models are useful when it is known that curvilinear effects are present in the true response function or as approximating functions (Taylor series expansion) to an unknown nonlinear relationship.

How To

- ✓ Run: **STATISTICS->REGRESSION -> POLYNOMIAL REGRESSION...**
- ✓ Select **DEPENDENT** (**RESPONSE**) variable and **INDEPENDENT** variable (**PREDICTOR**)**.**
- ✓ Enter the **DEGREE OF POLYNOMIAL** to fit (referred as *k* below)**.**
    - When the degree of a polynomial is equal to 1, the model is identical to the linear regression.
    - For lower degrees of *k*, the regression has a specific name: k = 2 – quadratic regression, k = 3 – cubic regression, k = 4 – quartic regression, k = 5 – quintic regression.
    - It is recommended to keep the degree of a polynomial as low as possible and avoid using high-order polynomials unless they can be justified for reasons outside the data (Montgomery, et al., 2013). High degrees may also risk a numerical overflow when values of the predictor variable are large.
    - As a general rule, k < 5 (Draper, Smith, 1998).
- ✓ Optionally, following charts can be included in the report:
    - Residuals versus predicted values plot (use the **PLOT RESIDUALS VS. FITTED** option);
    - Residuals versus order of observation plot (use the **PLOT RESIDUALS VS. ORDER** option);
    - Predicted values versus the observed values plot (**LINE FIT PLOT**).

Results

Report includes the regression statistics, analysis of variance (ANOVA) and tables with coefficients and residuals.

Regression Statistics

R^2 (Coefficient of determination, R-squared) - is the square of the sample correlation coefficient between the **Predictor** (independent variable) and **Response** (dependent variable).

Adjusted R2 (Adjusted R-squared) is a modification of R^2 that adjusts for the number of explanatory terms in a model.

See the **LINEAR REGRESSION** chapter for more details.

ANOVA Table

SOURCE OF VARIATION - the source of variation (term in the model). The **TOTAL** variance is partitioned into the variance, which can be explained by the independent variables (**REGRESSION**), and the variance, which is not explained by the independent variables (**ERROR**, sometimes called **RESIDUAL**). **SS (SUM OF SQUARES)** - the sum of squares for the term.

The line in the ANOVA table for the total gives the residual sum of squares corresponding to the mean function with the fewest parameters.

DF (DEGREES OF FREEDOM) - the number of observations for the corresponding model term. The **TOTAL** variance has $N - 1$ degrees of freedom. The **REGRESSION** degrees of freedom correspond to the number of coefficients estimated, including the intercept, minus 1.

MS (MEAN SQUARE) - an estimate of the variation accounted for by this term.

$$MS = SS/DF$$

F - the F-test value.

P-LEVEL - the significance level of the F-test. A value less than α shows that the model estimated by the regression procedure is significant.

Coefficients and Standard Errors Table

Regression coefficient (Beta), its standard error and confidence limits, the p-level and the risk ratio are displayed for each power of the predictor.

BETA – covariate regression coefficient estimate.

STANDARD ERROR – the standard error of the regression coefficient (Beta).

T-TEST – the t-statistics used in testing whether a given coefficient is significantly different from zero.

P-LEVEL - p-values for the null hypothesis that the coefficient is 0. Low p-value (< 0.05) allows the null hypothesis to be rejected and means that the covariate significantly improves the fit of the model.

LCL, UCL [BETA] – are the lower and upper 95% confidence intervals for the Beta, respectively. Default α level can be changed in the Preferences.

H0 (5%) - shows if null-hypothesis can be rejected/accepted at 5% level.

Residuals

PREDICTED values or fitted values are the values that the model predicts for each case using the regression equation.

RESIDUALS are differences between the observed values and the corresponding predicted values. Residuals represent the variance that is not explained by the model. The better the fit of the model, the smaller the values of residuals. Residuals are computed using the formula

$$e_i = Observed\ value\ -\ Predicted\ value = y_i - \widehat{y_i}.$$

Both the sum and the mean of the residuals are equal to zero.

Model

The polynomial regression model for a single predictor, *x*, is: $Y = c + a_1x + a_2x^2 + \cdots + a_kx^k + e$, where *Y* is the dependent variable, and *a*'s are the regression coefficients for the corresponding power of the predictor x^i, *c* is the constant or ***intercept***, and ***e*** is the error term reflected in the residuals. The regression function is linear in terms of the unknown parameters $c, a_1, \cdots, a_k$ because the powers of the predictor x, are treated as distinct independent variables x^i. For this reason, polynomial regression is considered as a form of a multiple linear regression, although it is used to fit a nonlinear (polynomial) model to the data. Unlike the linear regression model, extrapolation beyond the limits of data is dangerous and may produce meaningless results for high degree polynomials due to the problem of oscillation at the edges of data interval (known as Runge's phenomenon).

References

Draper, N. R., & Smith, H. (1998). Applied regression analysis. New York: Wiley.

Weisberg, S. (2013). Applied linear regression, 4th Ed. New York: Wiley.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2013). Introduction to linear regression analysis. Oxford: Wiley-Blackwell.