Statistics for the Sciences

Moving Beyond Linearity

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Polynomial Regression
- Local Regression
- Lab

Polynomial Regression

- Example (`caley.csv`): We will use the data to examine the regression of local species richness against regional species richness just for North America and at a sampling scale of 10% of the region. Although there was some evidence that both local and regional species richness were skewed, we will, like the original authors, analyze untransformed variables.
  - Response variable: `lspp10`
  - Predictor: `rspp10`

```
##      taxon lspp10 rspp10
## 1     AMPH      6      9
## 2    BIRDS    187    207
## 3   BUTTER    103    145
## 4     FISH     26     36
## 5  MAMMALS     66    117
## 6 REPTILES     59     80
## 7  ANGIOSP    130    172
## 8  GYMNOSP      9     11
```

Polynomial Regression

- SLR model fit
  - There seems a non-liner relationship between `lspp10` and `rspp10`

```
##
## Call:
## lm(formula = lspp10 ~ rspp10, data = caley)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.630  -6.147   1.496   5.718  23.194
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.79758    8.87011  -0.766    0.473
## rspp10       0.82417    0.07397  11.142 3.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.72 on 6 degrees of freedom
## Multiple R-squared:  0.9539, Adjusted R-squared:  0.9462
## F-statistic: 124.2 on 1 and 6 DF,  p-value: 3.116e-05
```

Polynomial Regression

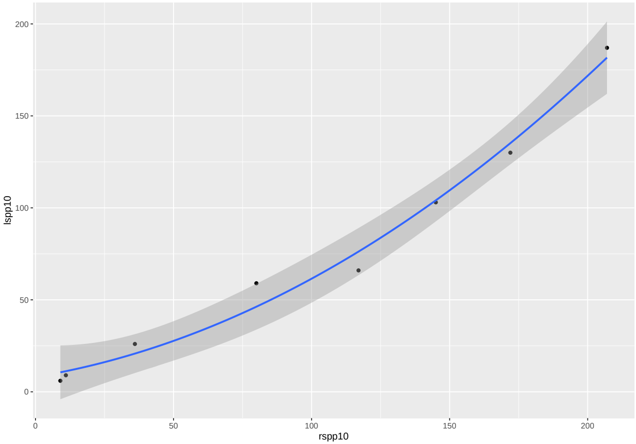
- The standard way to extend linear regression to nonlinear is to replace the standard linear model with a polynomial function

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \varepsilon_i$$

- Now we fit a Quadratic regression model `lspp10 ~ rspp10 + rspp10^2`

```
##
## Call:
## lm(formula = lspp10 ~ rspp10 + I(rspp10^2), data = caley)
##
## Residuals:
##       1       2       3       4       5       6       7       8
##  -4.595   5.332  -1.087   5.225 -10.227  12.739  -5.181  -2.207
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.1242187  6.7487986   1.204   0.2825
## rspp10      0.2488822  0.1700660   1.463   0.2032
## I(rspp10^2) 0.0028478  0.0008137   3.500   0.0173 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Polynomial Regression

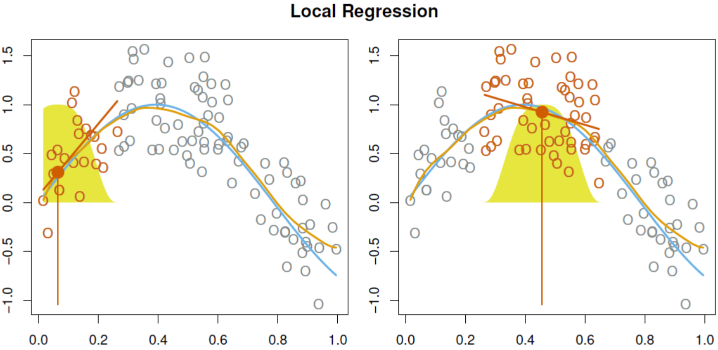


Polynomial Regression

- In performing a polynomial regression we must decide on the degree of the polynomial to use. One way to do this is by using hypothesis tests by comparing the models by performing an analysis of variance in order to test the null hypothesis that a model $\mathcal{M}_1$ is sufficient to explain the data against the alternative hypothesis that a more complex model $\mathcal{M}_2$ is required. Unfortunately, SPSS is not designed to compare models.
- For example, compare the following models
  - `y ~ x`
  - `y ~ x + x^2`
  - `y ~ x + x^2 + x^3`
  - `y ~ x + x^2 + x^3 + x^4`

Local Regression

- Local regression is a different approach for fitting flexible non-linear functions, which involves computing the fit at a target point x_0 using only the nearby observations.



Local Regression

- With a sliding weight function, we fit separate linear fits over the range of X by weighted least squares.
  - Weighted least squares seems to have been rarely applied in the biological literature, so we skip the topic in this course.
- Algorithm: Local Regression At $X = x_0$
  - 1. Gather the fraction $s = k/n$ of training points whose x_i are closest to x_0.
  - 2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood, so that the point furthest from x_0 has weight zero, and the closest has the highest weight. All but these k nearest neighbors get weight zero.
  - 3. Fit a weighted least squares regression of the y_i on the x_i using the aforementioned weights, by finding $\hat{\beta}_0$ and $\hat{\beta}_0$ that minimize

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2$$

  - 4. The fitted value at x_0 is given by $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

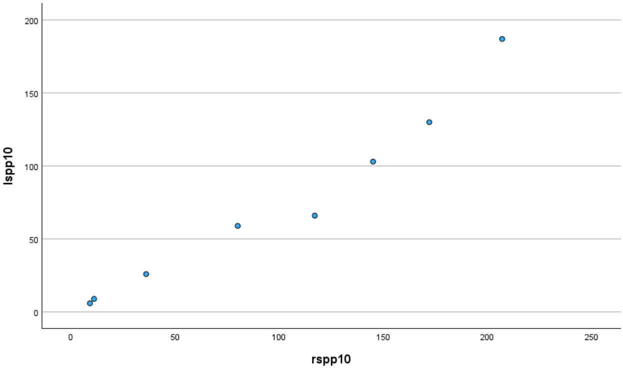
Lab

- Consider data `caley.csv`. After importing the data, conver the predictor and response to `scale` measure

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	taxon	String	8	0		None	None	8	Left	Nominal	Input
2	lspp10	Numeric	3	0		None	None	8	Right	Scale	Input
3	rspp10	Numeric	3	0		None	None	8	Right	Scale	Input
4											

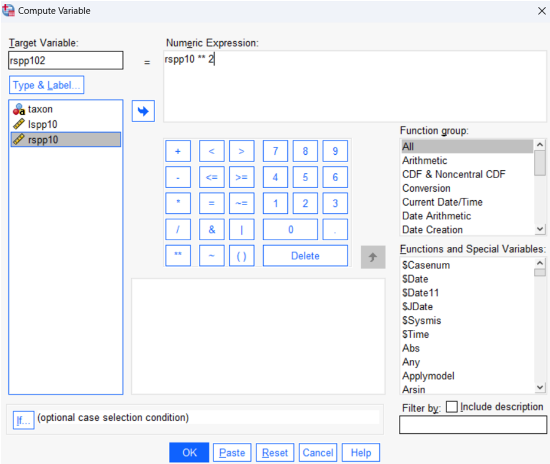
Lab

- Check the scatter plot of `rspp10`(x-axis) versus `lspp10` (y-axis).



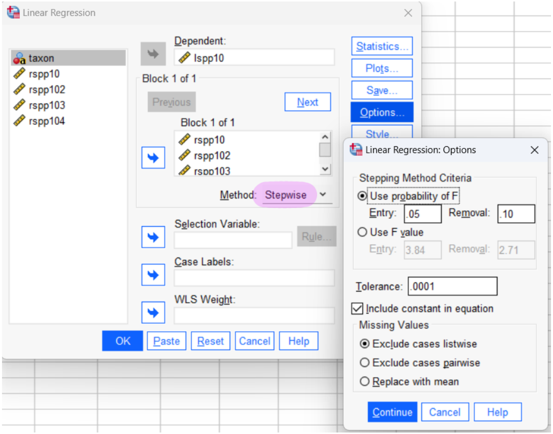
Lab

- Add three more columns of data: `rspp102=rspp10^2`, `rspp103=rspp10^3`, and `rspp104=rspp10^4`



Lab

- Then we use stepwise regression to check which predictors among `rspp10`, `rspp102`, `rspp103`, `rspp104` are significant:



Lab

- We can see that the final model is `lspp10 ~ rspp102`

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	15.746	4.683		3.362	.015
	rspp102	.004	.000	.990	17.560	<.001

a. Dependent Variable: lspp10

Excluded Variables[a]

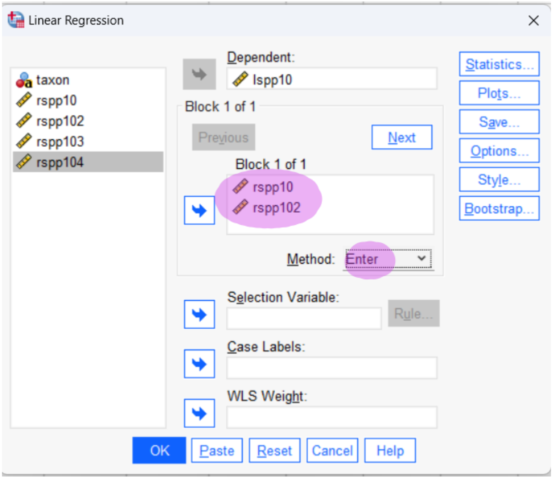
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	rspp10	.295[b]	1.463	.203	.548	.066
	rspp103	-.219[b]	-.649	.545	-.279	.031
	rspp104	-.094[b]	-.463	.663	-.203	.089

a. Dependent Variable: lspp10

b. Predictors in the Model: (Constant), rspp102

Lab

- If we want to follow the hierarchical principle, we can refit the model `lspp10 ~ rspp10 + rspp102`



Lab

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27806.880	2	13903.440	184.582	<.001[b]
	Residual	376.620	5	75.324		
	Total	28183.500	7			

a. Dependent Variable: lspp10

b. Predictors: (Constant), rspp102, rspp10

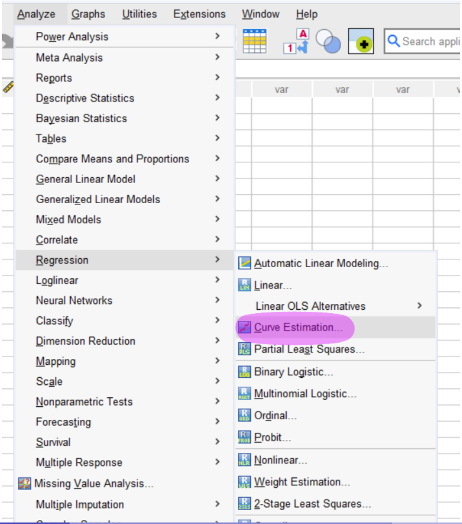
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.124	6.749		1.204	.283
	rspp10	.249	.170	.295	1.463	.203
	rspp102	.003	.001	.705	3.500	.017

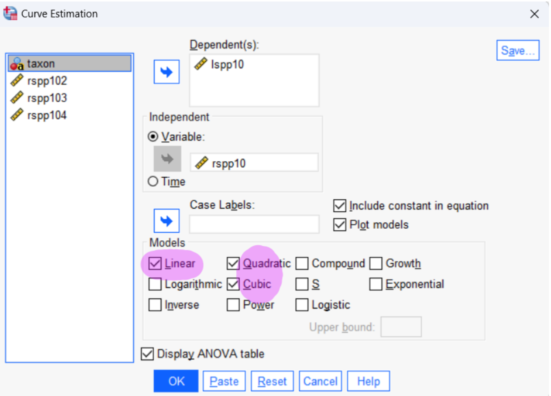
a. Dependent Variable: lspp10

Lab

- Another way is we can manually **compare the Adjusted R-squares** of the models and SPSS can give polynomial model up to degree 3.
  - The final model is different from stepwise regression.



Lab


Curve Estimation
taxon
rspp102
rspp103
rspp104
Dependent(s):
lspp10
Save...
Independent
Variable:
rspp10
Time
Case Labels:
Include constant in equation
Plot models
Models
Linear
Quadratic
Compound
Growth
Logarithmic
Cubic
S
Exponential
Inverse
Power
Logistic
Upper bound:
Display ANOVA table
OK
Paste
Reset
Cancel
Help

Lab

Linear

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.977	.954	.946	14.715

The independent variable is rspp10.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	26884.243	1	26884.243	124.152	<.001
Residual	1299.257	6	216.543		
Total	28183.500	7			

The independent variable is rspp10.

Lab

Quadratic

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.993	.987	.981	8.679

The independent variable is rspp10.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	27806.880	2	13903.440	184.582	<.001
Residual	376.620	5	75.324		
Total	28183.500	7			

The independent variable is rspp10.

Lab

Cubic

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.997	.995	.991	6.102

The independent variable is rspp10.

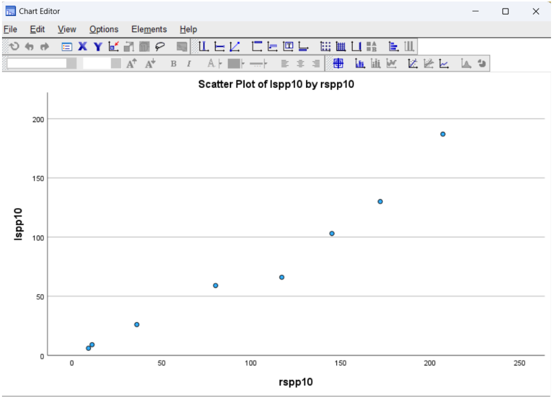
ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	28034.577	3	9344.859	250.998	<.001
Residual	148.923	4	37.231		
Total	28183.500	7			

The independent variable is rspp10.

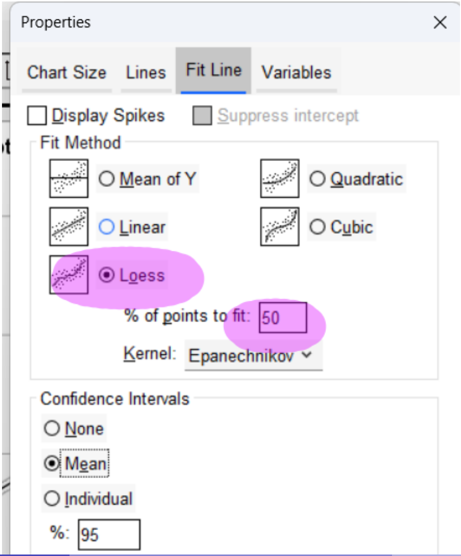
Lab

- LOESS is available from the Fit Line tab of the Properties panel when you edit a scatterplot in the chart editor.
- Create a scatter plot using `Graphs` $\rightarrow$ `Chart Builder`.
- Double-click on the scatter plot to open the Chart Editor.



Lab

- In the Chart Editor, go to the `Elements` tab → `Fit line at Total`
  - Adjust the Smoothing Parameter if necessary



License