Linear Models 3: Non-linearity Lab

Dr. Luisa Barbanti and Dr. Matteo Tanadini

Applied Machine Learning and Predictive Modelling 1, FS25 (HSLU)

Contents

1 Load packages

```
# Linear Models 3: Non-linearity Lab
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline
from statsmodels.formula.api import ols
from statsmodels.gam.api import GLMGam, BSplines
from statsmodels.stats.anova import anova_lm
from statsmodels.tools.tools import add_constant
from statsmodels.graphics.gofplots import qqplot
from statsmodels.stats.outliers_influence import variance_inflation_factor
import statsmodels.formula.api as smf
from patsy import dmatrix, bs
import statsmodels.api as sm
from pygam import GAM, s
```

2 Getting data

```
## Load data
d_trees = pd.read_csv("../../Datasets/TreesChamagne2017_Lab_modified.csv",
                      sep = ';', decimal = ',')

## Rename variables because the "." causes problem in python
d_trees.rename(columns = {'growth.rate': 'growth_rate'}, inplace = True)
d_trees.rename(columns = {'diversity.site': 'diversity_site'}, inplace = True)
d_trees.rename(columns = {'density.site': 'density_site'}, inplace = True)
```

```
# Inspect data
print(d_trees.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 557 entries, 0 to 556
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   growth_rate         557 non-null    float64
 1   species             557 non-null    object
 2   site                557 non-null    int64
 3   Density.tree.Class  557 non-null    object
 4   age                 557 non-null    float64
 5   size                557 non-null    float64
 6   density_site        557 non-null    float64
 7   density.tree        557 non-null    float64
 8   diversity.tree      557 non-null    float64
```

```
 9   diversity_site      557 non-null    float64
 10  sp.richness         557 non-null    int64
 11  SiteID              557 non-null    int64
dtypes: float64(7), int64(3), object(2)
memory usage: 52.3+ KB
None
```

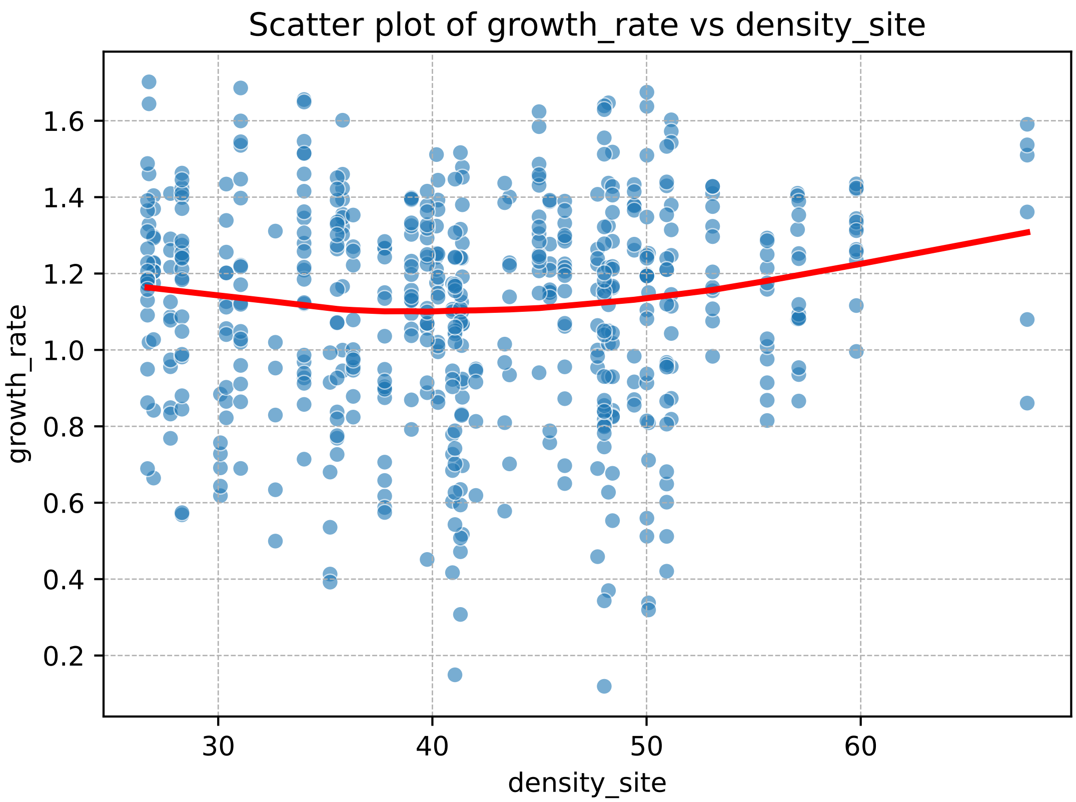
```
print(d_trees.head())
```

```
   growth_rate species  site  ... diversity_site  sp.richness  SiteID
0     0.701705   Beech     1  ...       1.279284            1       1
1     1.138995   Beech     1  ...       1.279284            1       1
2     1.394101   Beech    12  ...       2.272922            2      12
3     0.999519  Spruce    12  ...       2.272922            2      12
4     1.354924  Spruce    12  ...       2.272922            2      12

[5 rows x 12 columns]
```

3 Polynomials

3.1 Graphical analysis

```
plt.clf()
## Plotting the data
plt.grid(True, which = 'both', linestyle = '--', linewidth = 0.5)
sns.scatterplot(data = d_trees, x = 'density_site', y = 'growth_rate', alpha = 0.6)
## Add smoother
sns.regplot(data = d_trees, x = 'density_site', y = 'growth_rate',
            lowess = True,
            ## You could set 'scatter' equal to true, and then you would not
            ## need the sns.scatterplot function. However, the points would
            ## be red as the smoother and not a bit transparent.
            scatter = False, color = 'red')
plt.title('Scatter plot of growth_rate vs density_site')
```



3.2 Quadratic effect

```
## Linear model with linear effect for density_site
lm_trees_1 = ols('growth_rate ~ species + age + diversity_site + density_site',
                 data = d_trees).fit()
print(lm_trees_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            growth_rate   R-squared:                       0.183
Model:                            OLS   Adj. R-squared:                  0.174
Method:                 Least Squares   F-statistic:                     20.50
Date:                Tue, 25 Feb 2025   Prob (F-statistic):           1.03e-21
Time:                        10:10:58   Log-Likelihood:                -25.404
No. Observations:                 557   AIC:                             64.81
Df Residuals:                     550   BIC:                             95.07
Df Model:                           6
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          1.0483      0.078     13.480      0.000       0.896       1.201
species[T.Larch]  -0.3052      0.031     -9.812      0.000      -0.366      -0.244
species[T.Oak]    -0.1913      0.031     -6.237      0.000      -0.252      -0.131
```

```
species[T.Spruce]    -0.1008      0.031     -3.228      0.001      -0.162      -0.039
age                  -0.0004      0.001     -0.667      0.505      -0.001       0.001
diversity_site        0.0481      0.015      3.171      0.002       0.018       0.078
density_site          0.0026      0.001      1.973      0.049     1.2e-05       0.005
==============================================================================
Omnibus:                       34.295   Durbin-Watson:                   1.702
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               39.334
Skew:                          -0.602   Prob(JB):                     2.88e-09
Kurtosis:                       3.495   Cond. No.                         609.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

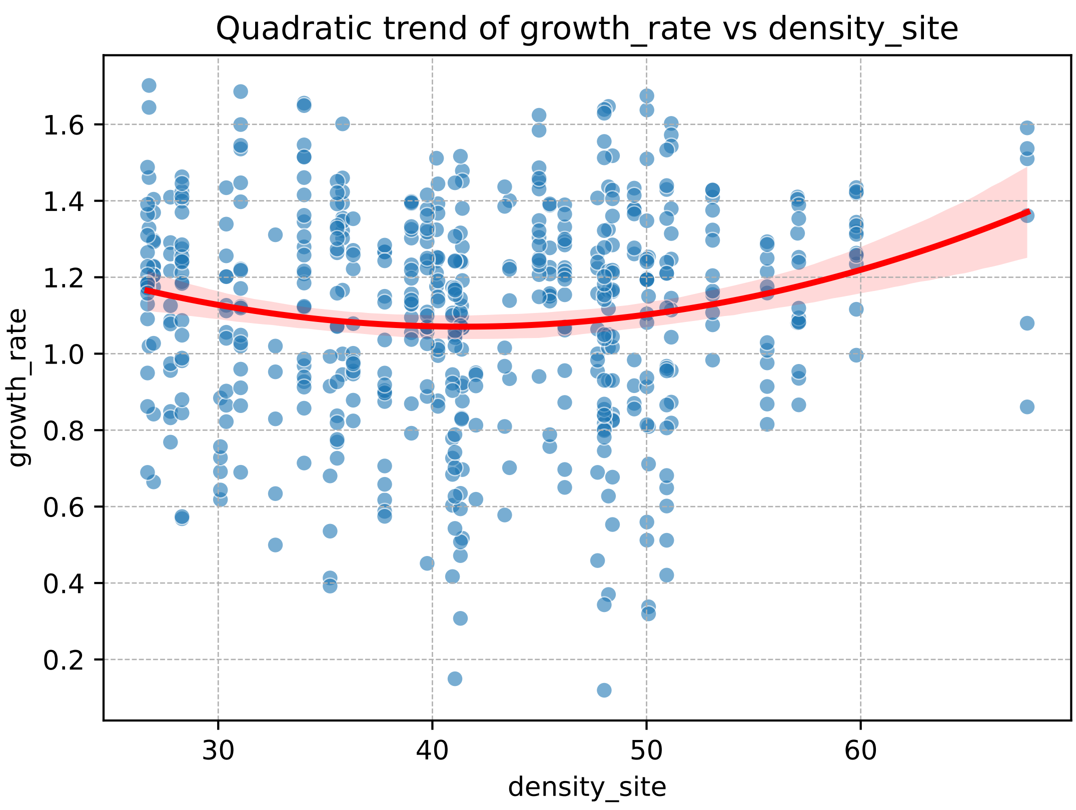
```
## Linear model with quadratic effect for density_site
lm_trees_2 = ols(
  'growth_rate ~ species + age + diversity_site + density_site + I(density_site**2)',
  data = d_trees).fit()
print(lm_trees_2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            growth_rate   R-squared:                       0.205
Model:                            OLS   Adj. R-squared:                  0.195
Method:                 Least Squares   F-statistic:                     20.25
Date:                Tue, 25 Feb 2025   Prob (F-statistic):           3.10e-24
Time:                        10:10:58   Log-Likelihood:                -17.645
No. Observations:                 557   AIC:                             51.29
Df Residuals:                     549   BIC:                             85.87
Df Model:                           7
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Intercept                 1.7359      0.191      9.102      0.000       1.361       2.110
species[T.Larch]         -0.2972      0.031     -9.659      0.000      -0.358      -0.237
species[T.Oak]           -0.1809      0.030     -5.953      0.000      -0.241      -0.121
species[T.Spruce]        -0.0974      0.031     -3.157      0.002      -0.158      -0.037
age                      -0.0004      0.001     -0.786      0.432      -0.002       0.001
diversity_site            0.0569      0.015      3.763      0.000       0.027       0.087
density_site             -0.0328      0.009     -3.605      0.000      -0.051      -0.015
I(density_site ** 2)      0.0004      0.000      3.938      0.000       0.000       0.001
==============================================================================
Omnibus:                       30.839   Durbin-Watson:                   1.745
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               34.738
Skew:                          -0.568   Prob(JB):                     2.86e-08
Kurtosis:                       3.457   Cond. No.                     3.54e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.54e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
## ANOVA to compare models
anova_results = anova_lm(lm_trees_1, lm_trees_2)
print(anova_results)
```

```
   df_resid        ssr  df_diff   ss_diff          F    Pr(>F)
0     550.0  35.727017      0.0       NaN        NaN       NaN
1     549.0  34.745348      1.0  0.981668  15.511022  0.000093
```

```
plt.clf()
## Plot with quadratic trend
plt.grid(True, which = 'both', linestyle = '--', linewidth = 0.5)
sns.scatterplot(data = d_trees, x = 'density_site', y = 'growth_rate', alpha = 0.6)
sns.regplot(data = d_trees, x = 'density_site', y = 'growth_rate', order = 2,
            scatter = False, color = "red")
plt.title('Quadratic trend of growth_rate vs density_site')
```

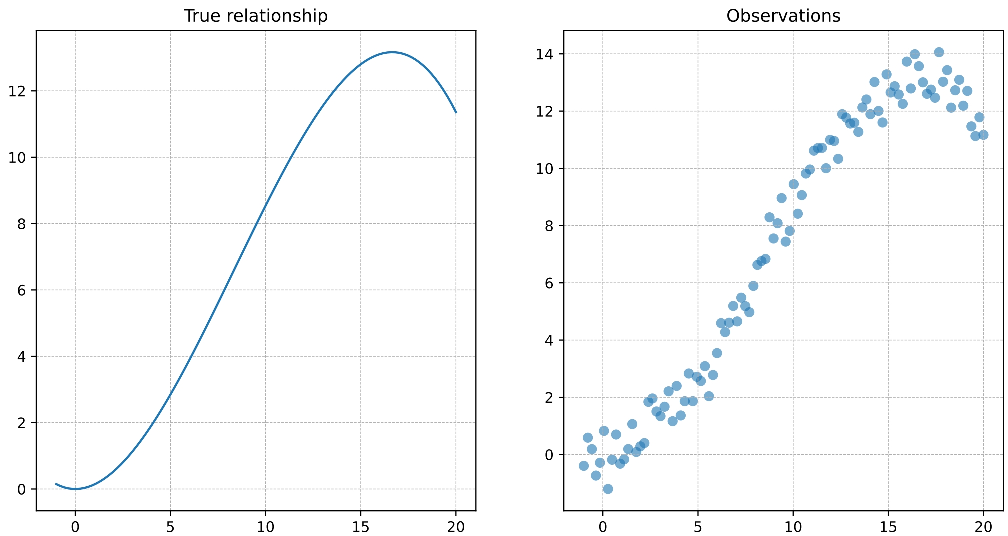


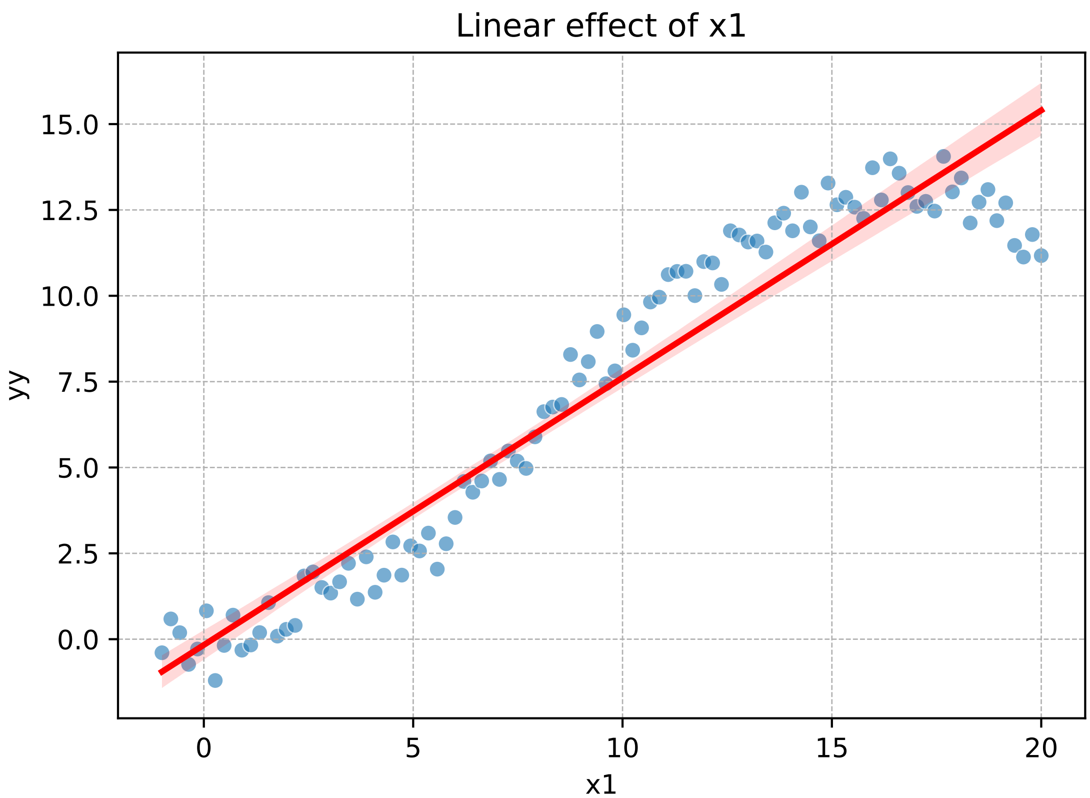

```
## Linear model with polynomial features
lm_trees_3 = ols(
  'growth_rate ~ species + age + diversity_site + np.power(density_site, 2)',
  data = d_trees).fit()
print(lm_trees_3.summary())
```

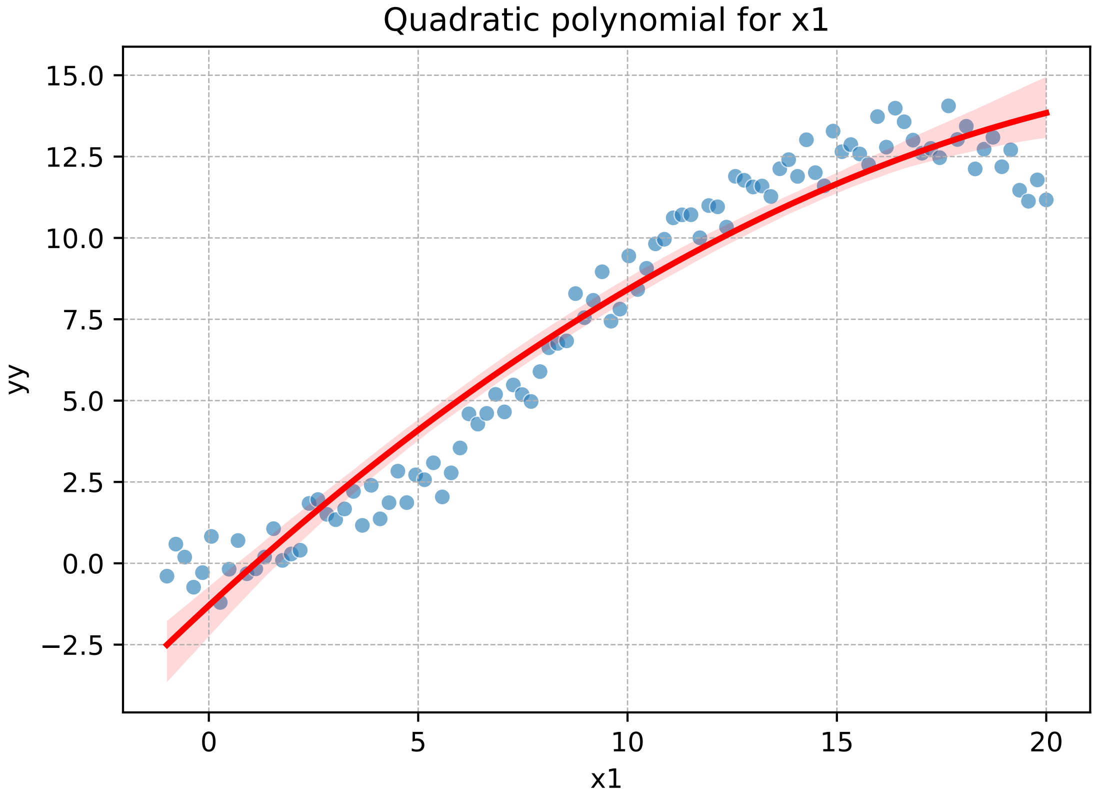
```
                            OLS Regression Results
```

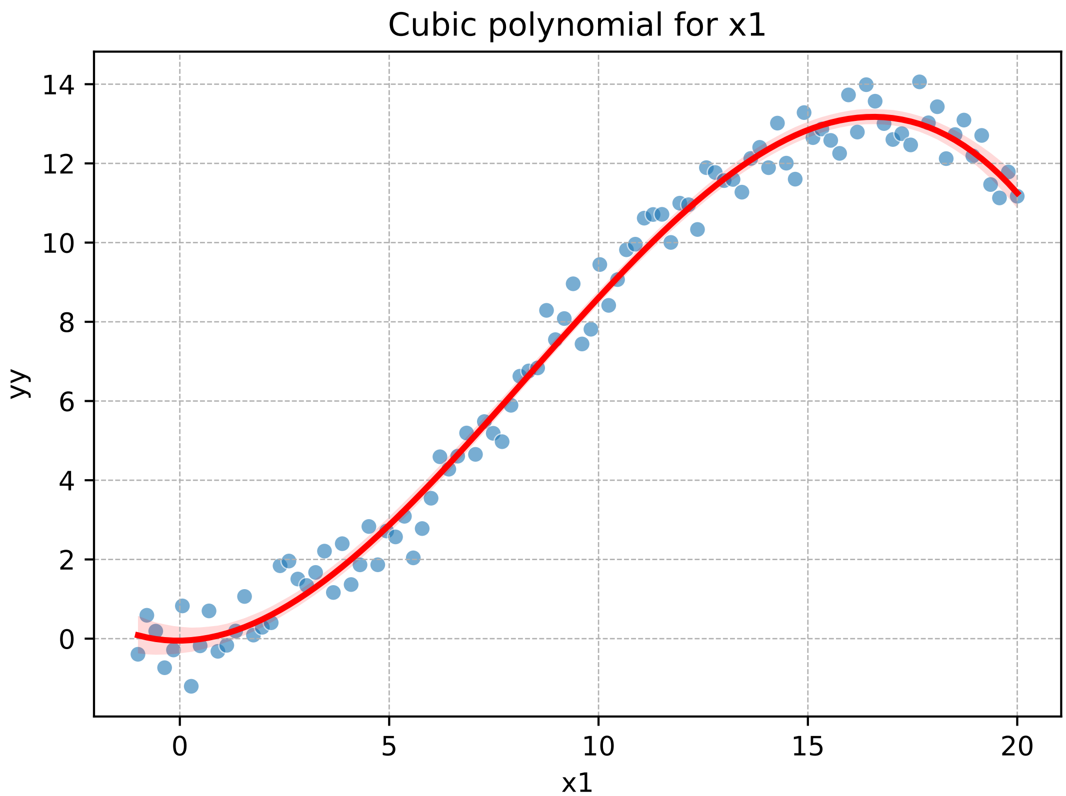
```
==============================================================================
Dep. Variable:            growth_rate   R-squared:                       0.186
Model:                            OLS   Adj. R-squared:                  0.178
Method:                 Least Squares   F-statistic:                     21.01
Date:                Tue, 25 Feb 2025   Prob (F-statistic):           3.13e-22
Time:                        10:10:59   Log-Likelihood:                -24.163
No. Observations:                 557   AIC:                             62.33
Df Residuals:                     550   BIC:                             92.58
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                     1.0882      0.065     16.810      0.000       0.961       1.215
species[T.Larch]             -0.3065      0.031     -9.888      0.000      -0.367      -0.246
species[T.Oak]               -0.1880      0.031     -6.133      0.000      -0.248      -0.128
species[T.Spruce]            -0.1037      0.031     -3.332      0.001      -0.165      -0.043
age                          -0.0005      0.001     -0.834      0.405      -0.002       0.001
diversity_site                0.0500      0.015      3.294      0.001       0.020       0.080
np.power(density_site, 2)  3.958e-05   1.57e-05      2.524      0.012    8.77e-06    7.04e-05
==============================================================================
Omnibus:                       34.837   Durbin-Watson:                   1.710
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               40.100
Skew:                          -0.606   Prob(JB):                     1.96e-09
Kurtosis:                       3.508   Cond. No.                     1.22e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.22e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

3.3 More complex non-linear relationships





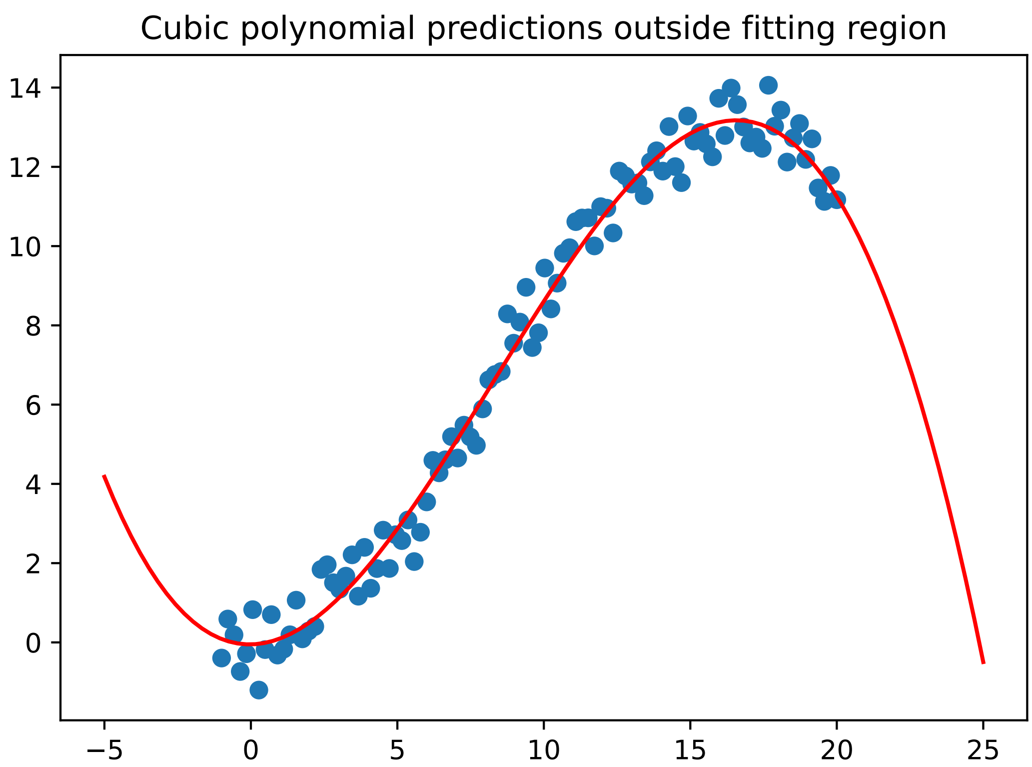

Quadratic polynomial for x1
yy
x1
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
0
5
10
15
20




```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     yy   R-squared:                       0.987
Model:                            OLS   Adj. R-squared:                  0.987
Method:                 Least Squares   F-statistic:                     2462.
Date:                Tue, 25 Feb 2025   Prob (F-statistic):           1.22e-90
Time:                        10:10:59   Log-Likelihood:                -84.287
No. Observations:                 100   AIC:                             176.6
Df Residuals:                      96   BIC:                             187.0
Df Model:                           3
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept            -0.0505      0.155     -0.326      0.745      -0.359       0.258
x1                    0.0131      0.073      0.179      0.858      -0.132       0.158
np.power(x1, 2)       0.1429      0.009     15.484      0.000       0.125       0.161
np.power(x1, 3)      -0.0058      0.000    -18.119      0.000      -0.006      -0.005
==============================================================================
Omnibus:                        2.410   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.300   Jarque-Bera (JB):                1.612
Skew:                           0.031   Prob(JB):                        0.447
Kurtosis:                       2.381   Cond. No.                     8.62e+03
==============================================================================

Notes:
```

```
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.62e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

3.4 Are polynomials the ultimate solution for modelling non-linear relationships?

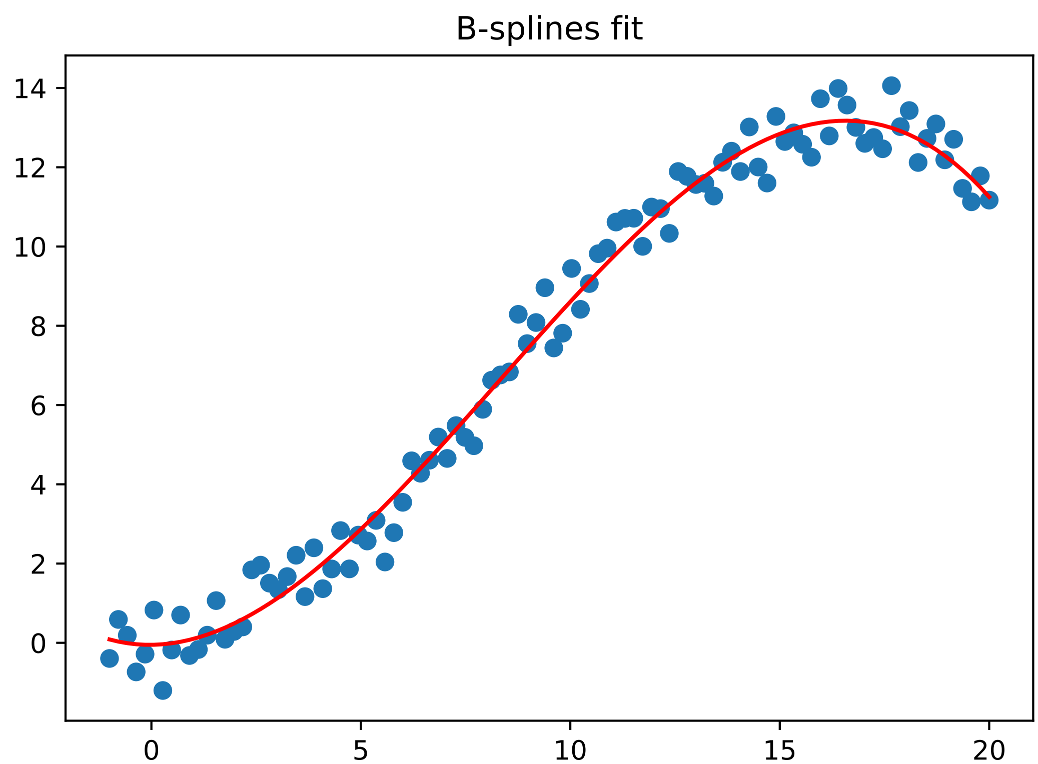


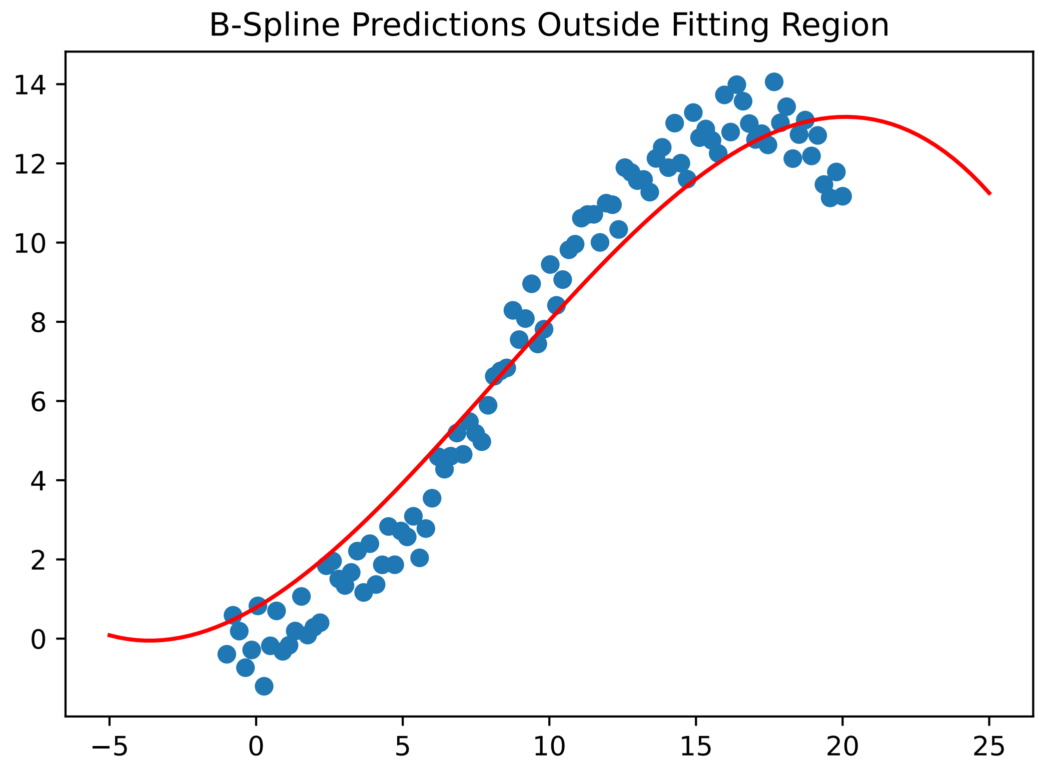
4 Regression splines

```python
# Generate B-spline basis with 3 degrees of freedom
X_spline = dmatrix(bs(sim_data['x1'], df = 3, include_intercept = False),
                   return_type = 'dataframe')


# Fit the linear model using regression splines
lm_regression_splines = sm.OLS(sim_data['yy'], sm.add_constant(X_spline)).fit()

# Print summary
print(lm_regression_splines.summary())
```

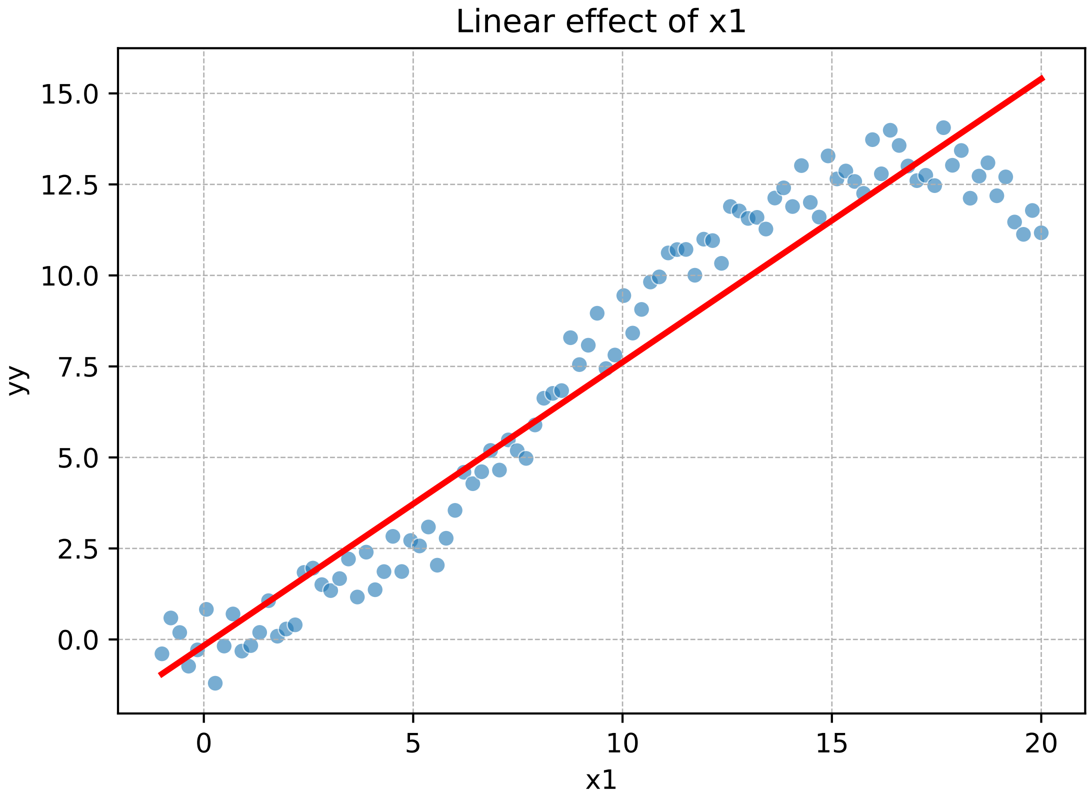
```
                            OLS Regression Results
==============================================================================
```

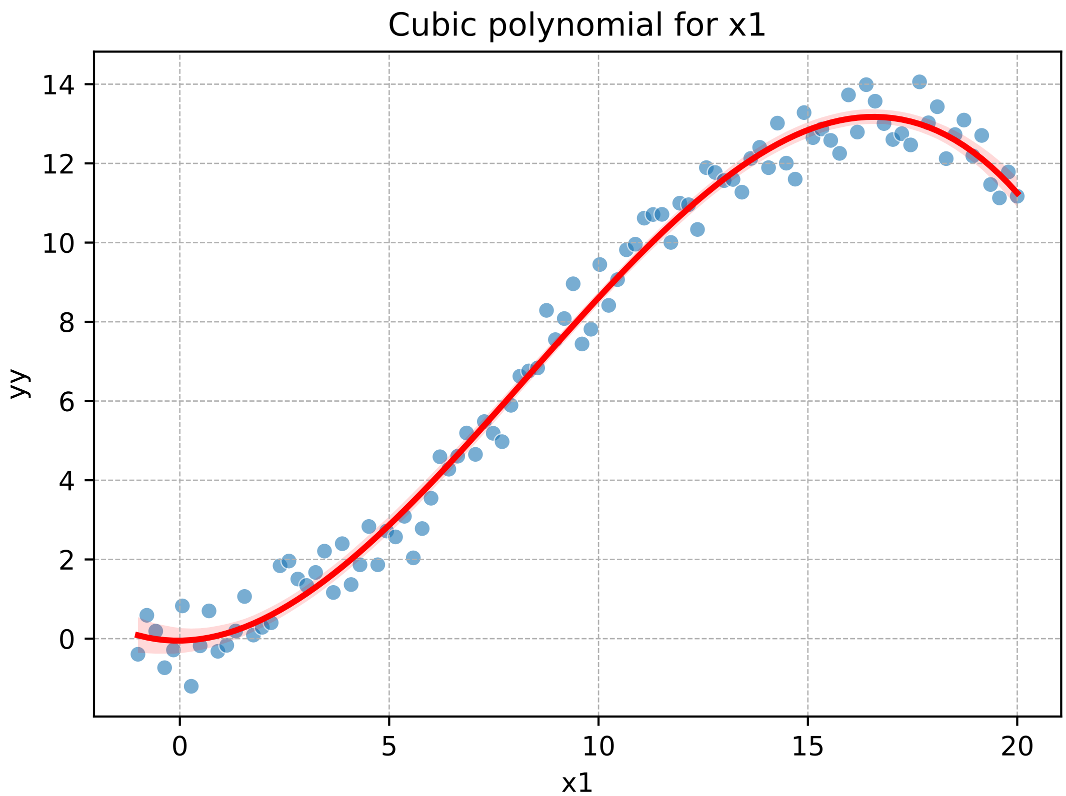
```
Dep. Variable:                     yy   R-squared:                       0.987
Model:                            OLS   Adj. R-squared:                  0.987
Method:                 Least Squares   F-statistic:                     2462.
Date:                Tue, 25 Feb 2025   Prob (F-statistic):           1.22e-90
Time:                        10:11:00   Log-Likelihood:                -84.287
No. Observations:                 100   AIC:                             176.6
Df Residuals:                      96   BIC:                             187.0
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0851      0.221      0.385      0.701      -0.354       0.524
x0            -2.0307      0.642     -3.165      0.002      -3.304      -0.757
x1            19.4937      0.413     47.158      0.000      18.673      20.314
x2            11.1659      0.347     32.173      0.000      10.477      11.855
==============================================================================
Omnibus:                        2.410   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.300   Jarque-Bera (JB):                1.612
Skew:                           0.031   Prob(JB):                        0.447
Kurtosis:                       2.381   Cond. No.                         14.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

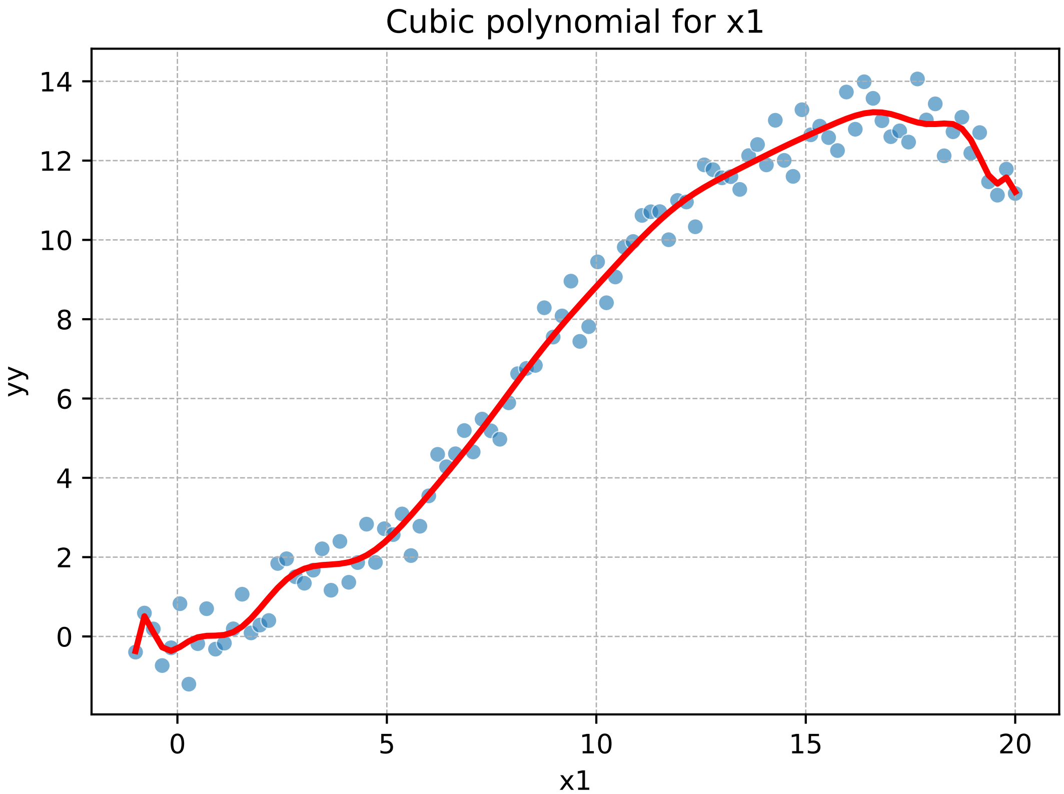



B-Spline Predictions Outside Fitting Region
14
12
10
8
6
4
2
0
−5
0
5
10
15
20
25

4.1 Degree of complexity, how much is enough?




Cubic polynomial for x1
yy
x1
14
12
10
8
6
4
2
0
0
5
10
15
20



5 Generalised Additive Models

```
## python doesn't seem to have a simple option
```