Linear Models 2: Tree growth Lab

Dr. Luisa Barbanti and Dr. Matteo Tanadini

Applied Machine Learning and Predictive Modelling 1, HS25 (HSLU)

Contents

1 Load package

```
## Load packages
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.stats.multicomp import pairwise_tukeyhsd
```

2 Getting data

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

```
# Rename variables because "." causes problems in python
d_trees.rename(columns = {'growth.rate': 'growth_rate'}, inplace = True)
d_trees.rename(columns = {'Density.tree.Class': 'Density_tree_Class'}, inplace = True)
```

```
## Inspect the 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())
```

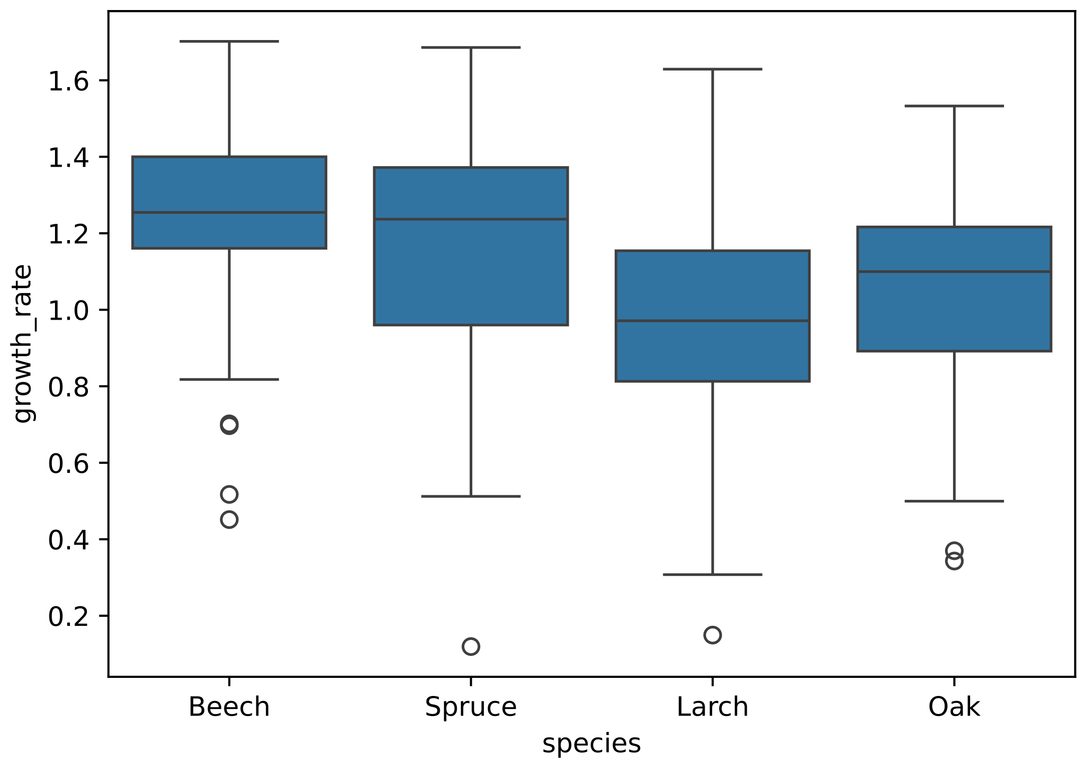
```
   growth_rate species  site  ... diversity.site  sp.richness  SiteID
0     0.701705   Beech     1  ...       1.279284            1       1
```

```
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]
```

```
## Clean figure object
plt.clf()

## Boxplot for growth rate by species
sns.boxplot(x = 'species', y = 'growth_rate', data = d_trees)
```



3 Fit linear models

```
## Fit a linear model: growth.rate ~ species
lm_trees_1 = smf.ols('growth_rate ~ species', data = d_trees).fit()
print(lm_trees_1.params)
```

```
Intercept            1.252760
species[T.Larch]    -0.299561
species[T.Oak]      -0.200879
```

```
species[T.Spruce]   -0.086747
dtype: float64
```

```
## Summary
print(lm_trees_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            growth_rate   R-squared:                       0.163
Model:                            OLS   Adj. R-squared:                  0.159
Method:                 Least Squares   F-statistic:                     35.99
Date:                Thu, 11 Sep 2025   Prob (F-statistic):           2.92e-21
Time:                        14:40:36   Log-Likelihood:                -31.948
No. Observations:                 557   AIC:                             71.90
Df Residuals:                     553   BIC:                             89.19
Df Model:                           3
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept             1.2528      0.022     58.046      0.000       1.210       1.295
species[T.Larch]     -0.2996      0.031     -9.708      0.000      -0.360      -0.239
species[T.Oak]       -0.2009      0.030     -6.593      0.000      -0.261      -0.141
species[T.Spruce]    -0.0867      0.031     -2.811      0.005      -0.147      -0.026
==============================================================================
Omnibus:                       31.045   Durbin-Watson:                   1.653
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               34.746
Skew:                          -0.583   Prob(JB):                     2.85e-08
Kurtosis:                       3.371   Cond. No.                         4.75
==============================================================================

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

```
## Fit a null model: growth.rate ~ 1
lm_trees_0 = smf.ols('growth_rate ~ 1', data = d_trees).fit()
print(lm_trees_0.params)
```

```
Intercept    1.106865
dtype: float64
```

```
## Summary
print(lm_trees_0.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            growth_rate   R-squared:                      -0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                       nan
Date:                Thu, 11 Sep 2025   Prob (F-statistic):                nan
Time:                        14:40:37   Log-Likelihood:                -81.623
No. Observations:                 557   AIC:                             165.2
```

```
Df Residuals:                     556   BIC:                             169.6
Df Model:                           0
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.1069      0.012     93.160      0.000       1.084       1.130
==============================================================================
Omnibus:                       26.924   Durbin-Watson:                   1.440
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.641
Skew:                          -0.560   Prob(JB):                     3.66e-07
Kurtosis:                       3.148   Cond. No.                         1.00
==============================================================================

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

```
## Compare the two models
anova_comparison = sm.stats.anova_lm(lm_trees_0, lm_trees_1)
print(anova_comparison)
```

```
   df_resid        ssr  df_diff  ss_diff          F        Pr(>F)
0     556.0  43.718384      0.0      NaN        NaN           NaN
1     553.0  36.576334      3.0  7.14205  35.993706  2.921792e-21
```

4 Contrasts

4.1 Oak vs Spruce

```
## Check whether Oak and Spruce differ in terms of growth rates
# Filter dataset to include only Oak and Spruce
d_trees_filtered = d_trees[d_trees['species'].isin(['Oak', 'Spruce'])]
tukey_results = pairwise_tukeyhsd(endog = d_trees_filtered['growth_rate'].astype(float),
                                  groups = d_trees_filtered['species'])

# Print summary
print(tukey_results)
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
==================================================
group1 group2 meandiff p-adj  lower  upper  reject
--------------------------------------------------
   Oak Spruce   0.1141 0.0003 0.0525 0.1758   True
--------------------------------------------------
```

5 Testing several variables

5.1 Testing categorical variables

```python
## Add Density.tree.Class and SiteID to the model
lm_trees_2 = smf.ols('growth_rate ~ species + Density_tree_Class + SiteID',
                     data = d_trees).fit()
print(lm_trees_2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            growth_rate   R-squared:                       0.167
Model:                            OLS   Adj. R-squared:                  0.158
Method:                 Least Squares   F-statistic:                     18.43
Date:                Thu, 11 Sep 2025   Prob (F-statistic):           1.46e-19
Time:                        14:40:37   Log-Likelihood:                -30.602
No. Observations:                 557   AIC:                             75.20
Df Residuals:                     550   BIC:                             105.5
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                       1.2252      0.039     31.182      0.000       1.148       1.302
species[T.Larch]               -0.3045      0.031     -9.721      0.000      -0.366      -0.243
species[T.Oak]                 -0.2031      0.032     -6.326      0.000      -0.266      -0.140
species[T.Spruce]              -0.0962      0.031     -3.062      0.002      -0.158      -0.034
Density_tree_Class[T.low]      -0.0250      0.028     -0.885      0.377      -0.081       0.031
Density_tree_Class[T.medium]    0.0046      0.027      0.171      0.865      -0.049       0.058
SiteID                          0.0009      0.001      1.113      0.266      -0.001       0.002
==============================================================================
Omnibus:                       31.557   Durbin-Watson:                   1.682
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               35.512
Skew:                          -0.582   Prob(JB):                     1.94e-08
Kurtosis:                       3.419   Cond. No.                         205.
==============================================================================

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

```python
##
## Check
print(lm_trees_2.model.formula)
```

```
growth_rate ~ species + Density_tree_Class + SiteID
```

```python
## Test the two newly added variables
anova_added_var = sm.stats.anova_lm(lm_trees_2, typ = 2)
anova_added_var_df = pd.DataFrame(anova_added_var)
print(anova_added_var_df)
```

```
                        sum_sq     df          F        PR(>F)
species               6.851124    3.0  34.506581  1.906822e-20
Density_tree_Class    0.086108    2.0   0.650544  5.221626e-01
SiteID                0.081923    1.0   1.237854  2.663711e-01
Residual             36.399993  550.0        NaN           NaN
```

```
## SiteID wasn't correctly coded. We recode it.

## Add a factor version of SiteID
d_trees['SiteID_fac'] = d_trees['SiteID'].astype('category')
```

```
## Update model to use SiteID.fac instead of SiteID
lm_trees_3 = smf.ols('growth_rate ~ species + Density_tree_Class + SiteID_fac',
                     data = d_trees).fit()
# print(lm_trees_3.summary())
```

```
## Test the variables
anova_added_var2 = sm.stats.anova_lm(lm_trees_3, typ = 2)
anova_added_var2_df = pd.DataFrame(anova_added_var2)
print(anova_added_var2_df)
```

```
                        sum_sq     df          F        PR(>F)
species               3.973384    3.0  22.869942  6.705132e-14
Density_tree_Class    0.255194    2.0   2.203263  1.114986e-01
SiteID_fac            7.120148   44.0   2.794230  3.389389e-08
Residual             29.361768  507.0        NaN           NaN
```

5.2 Testing continuous and categorical variables

```
## Add age to the model
lm_trees_4 = smf.ols('growth_rate ~ species + Density_tree_Class + SiteID_fac + age',
                     data = d_trees).fit()
# print(lm_trees_4.summary())
```

```
## Global F-test
anova_global = sm.stats.anova_lm(lm_trees_4)
print(anova_global)
```

```
                       df     sum_sq   mean_sq          F        PR(>F)
species               3.0   7.142050  2.380683  41.095496  8.822530e-24
Density_tree_Class    2.0   0.094418  0.047209   0.814924  4.432532e-01
SiteID_fac           44.0   7.120148  0.161822   2.793373  3.439666e-08
age                   1.0   0.048928  0.048928   0.844590  3.585251e-01
Residual            506.0  29.312841  0.057931        NaN           NaN
```

```
## Check coefficient for age
# Get summary output as text
lm_trees_4_summary_text = lm_trees_4.summary().as_text()

# Convert to list of lines
```

```
summary_lines = lm_trees_4_summary_text.split("\n")

# Extract specific lines (equivalent to R's c(10,11,62))
selected_lines = [summary_lines[i] for i in [12, 13, 64]]

# Print the selected lines
for line in selected_lines:
    print(line)
```

```
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
age                             0.0015      0.002      0.919      0.359      -0.002       0.005
```

6 Appendix

6.1 Testing all predictors in a model

```
## Compare two models
anova_comparison2 = sm.stats.anova_lm(lm_trees_0, lm_trees_4)
print(anova_comparison2)
```

```
   df_resid        ssr  df_diff    ss_diff         F        Pr(>F)
0     556.0  43.718384      0.0        NaN       NaN           NaN
1     506.0  29.312841     50.0  14.405543  4.973387  9.535671e-22
```

6.2 Sequential sum of squares

```
## Sequential sum of squares
lm_trees_4.model.formula
```

```
'growth_rate ~ species + Density_tree_Class + SiteID_fac + age'
```

```
print(sm.stats.anova_lm(lm_trees_4))
```

```
                       df     sum_sq   mean_sq          F        PR(>F)
species               3.0   7.142050  2.380683  41.095496  8.822530e-24
Density_tree_Class    2.0   0.094418  0.047209   0.814924  4.432532e-01
SiteID_fac           44.0   7.120148  0.161822   2.793373  3.439666e-08
age                   1.0   0.048928  0.048928   0.844590  3.585251e-01
Residual            506.0  29.312841  0.057931        NaN           NaN
```

```
##
## Let's move *species* at the end
lm_trees_4_again = smf.ols('growth_rate ~ Density_tree_Class + SiteID_fac + age + species',
                     data = d_trees).fit()
print(sm.stats.anova_lm(lm_trees_4_again))
```

```
                      df     sum_sq   mean_sq          F        PR(>F)
Density_tree_Class   2.0   0.424338  0.212169   3.662472  2.635188e-02
SiteID_fac          44.0   9.958894  0.226338   3.907068  4.092440e-14
species              3.0   3.973384  1.324461  22.862932  6.792075e-14
age                  1.0   0.048928  0.048928   0.844590  3.585251e-01
Residual           506.0  29.312841  0.057931        NaN           NaN
```

6.3 Testing all pairwise comparisons

```
# Tukey HSD test for species (i.e., testing all pairwise comparisons)
tukey_species = pairwise_tukeyhsd(endog = d_trees['growth_rate'].astype(float),
                                  groups = d_trees['species'],
                                  alpha = 0.05)
print(tukey_species)
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower   upper  reject
----------------------------------------------------
 Beech  Larch  -0.2996    0.0 -0.3791   -0.22   True
 Beech    Oak  -0.2009    0.0 -0.2794 -0.1224   True
 Beech Spruce  -0.0867 0.0262 -0.1663 -0.0072   True
 Larch    Oak   0.0987 0.0078  0.0193  0.1781   True
 Larch Spruce   0.2128    0.0  0.1324  0.2932   True
   Oak Spruce   0.1141 0.0013  0.0348  0.1935   True
----------------------------------------------------
```

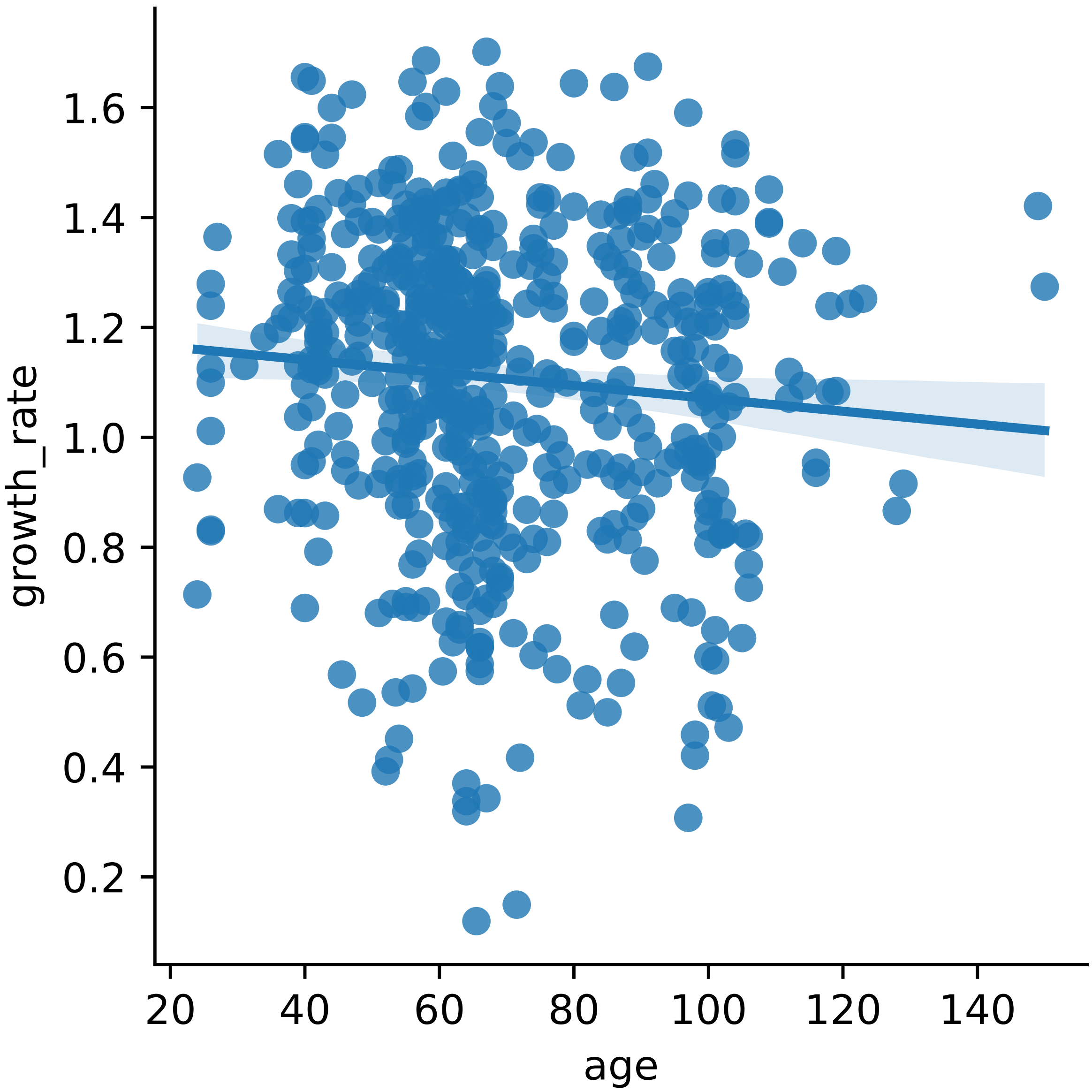
```
## Unfortunately, the plot doesn't seem to be as straightforward as in R
```

6.4 Principle of marginality

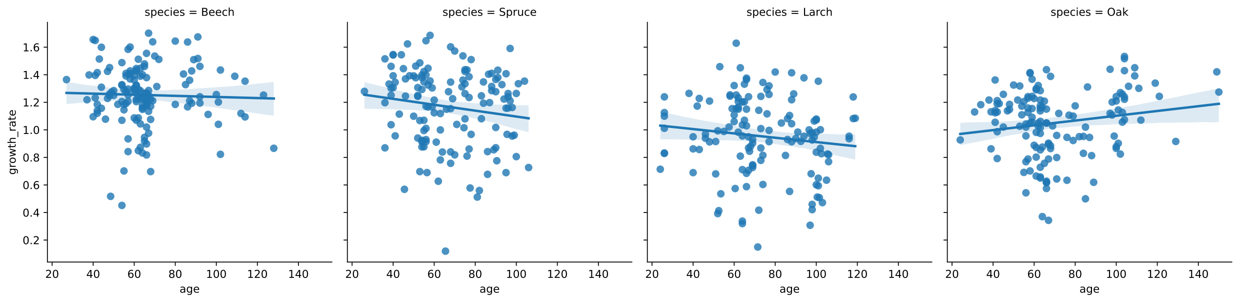
```
# Add interaction between age and species
lm_trees_5 = smf.ols('growth_rate ~ species + Density_tree_Class + SiteID_fac + age + age:species',
                     data = d_trees).fit()
# print(lm_trees_5.summary())
print(sm.stats.anova_lm(lm_trees_5))
```

```
                      df     sum_sq   mean_sq          F        PR(>F)
species              3.0   7.142050  2.380683  42.765326  1.248640e-24
Density_tree_Class   2.0   0.094418  0.047209   0.848036  4.288664e-01
SiteID_fac          44.0   7.120148  0.161822   2.906876  9.061162e-09
age                  1.0   0.048928  0.048928   0.878909  3.489507e-01
age:species          3.0   1.311566  0.437189   7.853422  3.946487e-05
Residual           503.0  28.001275  0.055669        NaN           NaN
```

```
plt.clf()
## Visualise age effect over all data
g = sns.lmplot(x = 'age', y = 'growth_rate', data = d_trees,
               height = 4, aspect = 1, ci = 95)
```




```
plt.clf()
## Visualise age effect for each species
g = sns.lmplot(x = 'age', y = 'growth_rate', data = d_trees, col = 'species',
               height = 4, aspect = 1, ci = 95)
```



6.5 Comparing F-tests and t-tests for categorical variables

```
print(lm_trees_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            growth_rate   R-squared:                       0.163
Model:                            OLS   Adj. R-squared:                  0.159
Method:                 Least Squares   F-statistic:                     35.99
Date:                Thu, 11 Sep 2025   Prob (F-statistic):           2.92e-21
Time:                        14:40:39   Log-Likelihood:                -31.948
No. Observations:                 557   AIC:                             71.90
Df Residuals:                     553   BIC:                             89.19
Df Model:                           3
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept             1.2528      0.022     58.046      0.000       1.210       1.295
species[T.Larch]     -0.2996      0.031     -9.708      0.000      -0.360      -0.239
species[T.Oak]       -0.2009      0.030     -6.593      0.000      -0.261      -0.141
species[T.Spruce]    -0.0867      0.031     -2.811      0.005      -0.147      -0.026
==============================================================================
Omnibus:                       31.045   Durbin-Watson:                   1.653
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               34.746
Skew:                          -0.583   Prob(JB):                     2.85e-08
Kurtosis:                       3.371   Cond. No.                         4.75
==============================================================================

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

```
print(sm.stats.anova_lm(lm_trees_1))
```

```
             df     sum_sq   mean_sq          F        PR(>F)
species     3.0   7.142050  2.380683  35.993706  2.921792e-21
Residual  553.0  36.576334  0.066142        NaN           NaN
```

```python
## Relevel *species*
d_trees['species_relevelled'] = pd.Categorical(d_trees['species'],
                                               categories = ['Oak', 'Beech', 'Spruce', 'Larch'],
                                               ordered = True)

lm_trees_1_relevelled = smf.ols('growth_rate ~ species_relevelled', data = d_trees).fit()
print(lm_trees_1_relevelled.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            growth_rate   R-squared:                       0.163
Model:                            OLS   Adj. R-squared:                  0.159
Method:                 Least Squares   F-statistic:                     35.99
Date:                Thu, 11 Sep 2025   Prob (F-statistic):           2.92e-21
Time:                        14:40:39   Log-Likelihood:                -31.948
No. Observations:                 557   AIC:                             71.90
Df Residuals:                     553   BIC:                             89.19
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                        1.0519      0.022     48.910      0.000       1.010       1.094
species_relevelled[T.Beech]      0.2009      0.030      6.593      0.000       0.141       0.261
species_relevelled[T.Spruce]     0.1141      0.031      3.705      0.000       0.054       0.175
```

```
species_relevelled[T.Larch]     -0.0987      0.031     -3.204      0.001      -0.159      -0.038
==============================================================================
Omnibus:                       31.045   Durbin-Watson:                   1.653
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               34.746
Skew:                          -0.583   Prob(JB):                     2.85e-08
Kurtosis:                       3.371   Cond. No.                         4.74
==============================================================================

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

```
print(sm.stats.anova_lm(lm_trees_1_relevelled))
```

```
                       df     sum_sq   mean_sq          F        PR(>F)
species_relevelled    3.0   7.142050  2.380683  35.993706  2.921792e-21
Residual            553.0  36.576334  0.066142        NaN           NaN
```