Complex Social Networks - Lab 4

Afonso Ramos
Aymeric Tavernier

November 11, 2019

1 Introduction

The objective of this fourth lab project is to practice on the fit of a non-linear function to data using collections of syntactic dependency trees from different languages. Just as a reminder, in this lab

n is defined as the number of vertices of a tree.

$< k^2 >$ is defined as its degree 2^{nd} moment and $< d >$ as the mean length of its edges.

From these two options we've selected the scaling of $< k^2 >$ as a function of n for the provided 10 different languages with all the following available models:

- $f(n) = (n/2)^{\text{b}}$ (model 1). This models is obtained applying the condition $f(2) = 1$ (satisfied both by $< d >$ and $< k^2 >$) to a more general model, i.e. $f(n) = an^{\text{b}}$. This leads to $a = 1/2^{\text{b}}$ and finally $f(n) = (n/2)^{\text{b}}$. The motivation of this model is that $< k^2 >=< d >= 1$ when $n = 2$.
- $f(n) = an^{\text{b}}$ (model 2), a power-law model.
- $f(n) = ae^{\text{cn}}$ (model 3), an exponential model.
- $f(n) = a\log(n)$ (model 4), a logarithmic model.

Furthermore, we were able to implement all the advanced models, accounting an additive term of the random linear arrangement model, being able to be generalised to, as detailed below:

- $f(n) = (n/2)^{\text{b}} + d$ (model 1+).
- $f(n) = an^{\text{b}} + d$ (model 2+).
- $f(n) = ae^{\text{cn}} + d$ (model 3+).
- $f(n) = a\log(n) + d$ (model 4+).

2 Results

Firstly, we started by generating the summary table for all the languages

Language	N	mu_n	sigma_n	mu_x	sigma_x
Arabic	4108	26.957644	426.28740	4.160443	521.33694
Basque	2933	11.335493	42.60344	4.143336	52.91292
Catalan	15053	25.571713	185.45671	4.961791	425.44742
Chinese	54238	6.248885	10.95861	3.218085	10.32232
Czech	25037	16.427647	114.94749	4.292722	148.94321
English	18779	24.046222	125.95388	5.170150	356.94974
Greek	2951	22.820400	206.76883	4.599747	333.13683
Hungarian	6424	21.659869	157.89069	5.955709	249.52997
Italian	4144	18.406612	178.06560	4.340449	199.22690
Turkish	6030	11.101658	68.57723	3.759063	54.78612

Figure 1: Summary table.

As instructed, we started by generating the preliminary plots, in this case, as a demonstration, for Catalan.

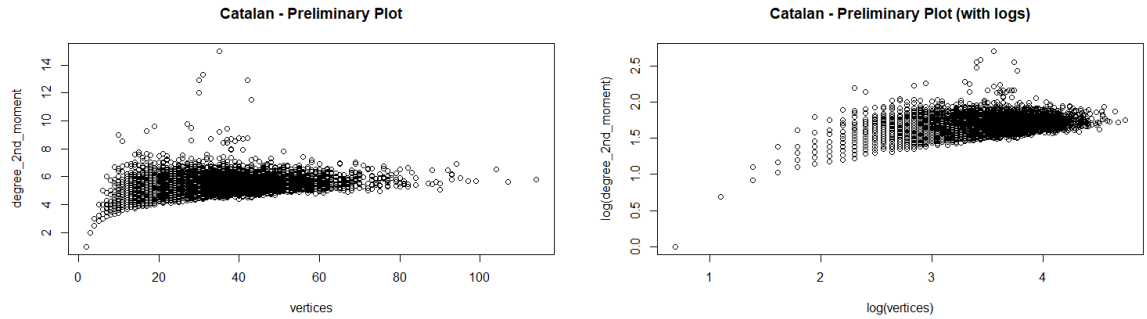

Figure 2: Preliminary Plots (Part 1).

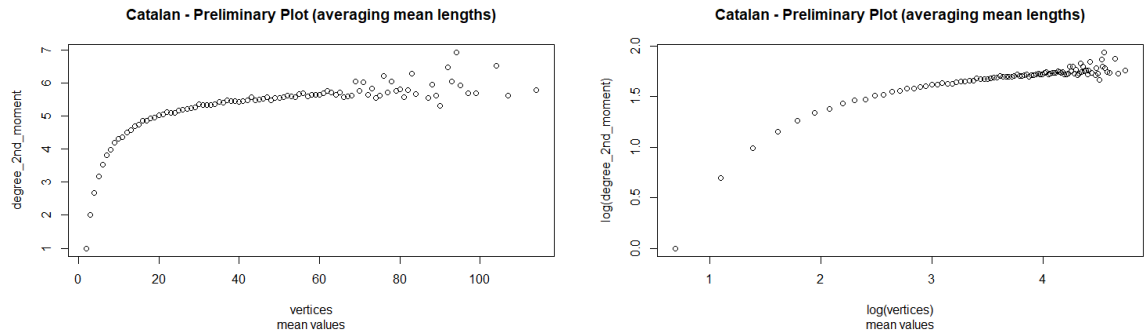

Figure 3: Preliminary Plots (Part 2).

2.1 Model Selection

Language	0	1	2	3	4	1+	2+	3+	4+
Arabic	8.024692	1.371610	0.7214354	0.7945320	0.8244184	0.7381814	0.7245119	0.7979202	0.7101210
Basque	1.568439	1.266747	0.4761478	0.6996916	0.5063874	0.5516400	0.4822136	0.7086052	0.3961709
Catalan	5.649863	1.372009	0.3889786	0.6054668	0.5561838	0.4593102	0.3910643	0.6087132	0.3293115
Chinese	1.066010	1.270552	0.4622285	0.6547797	0.5885212	0.5284227	0.4681170	0.6631212	0.3984868
Czech	3.229791	6.323645	6.3528770	6.5038957	6.6111555	6.3603044	6.3530764	6.5420421	6.5820075
English	4.635478	1.492730	0.5339175	0.7337273	0.6453667	0.5904673	0.5370490	0.7380307	0.4908780
Greek	5.342004	1.395595	0.6555272	0.7983149	0.7348555	0.6950942	0.6594644	0.8031097	0.6245618
Hungarian	3.815772	1.784165	0.9724685	1.1991892	0.9285518	1.0382917	0.9786824	1.2068519	0.9228295
Italian	4.330456	1.364508	0.5001893	0.7095665	0.5896824	0.5569172	0.5032300	0.7138800	0.4537302
Turkish	2.246577	1.228449	0.3919239	0.5622000	0.5877715	0.4487728	0.3955361	0.5673817	0.3422016

Figure 4: S Table as per generated by our R Scripts.

Language	0	1	2	3	4	1+	2+	3+	4+
Arabic	28770.01	419.3776	266.16540	289.32792	297.20256	271.67264	268.16540	291.32792	262.37162
Basque	10965.67	142.0407	60.81138	93.14394	65.02066	73.17345	62.81138	95.14394	45.36524
Catalan	94853.05	336.1559	95.12272	180.07799	162.79290	127.03339	97.12272	182.07799	63.15084
Chinese	160856.81	142.2926	58.31919	87.57131	77.64678	69.56151	60.31919	89.57131	45.85485
Czech	129761.55	577.3235	579.11787	583.25274	585.14894	579.32352	580.09415	585.25274	585.35391
English	110898.68	323.2342	143.26765	199.21736	175.65061	160.98609	145.26765	201.21736	128.47561
Greek	18265.97	304.3828	175.39555	209.29044	194.06157	185.47605	177.39555	211.29044	167.07255
Hungarian	35437.83	326.6518	229.32029	263.26954	220.85290	239.93038	231.32029	265.26954	220.83259
Italian	23909.66	297.0486	127.42496	186.86845	154.42451	145.68805	129.42496	188.86845	110.85272
Turkish	26875.86	188.2059	58.94067	100.07073	104.16857	74.38188	60.94067	102.07073	43.47457

Figure 5: AIC Table as per generated by our R Scripts.

Language	0	1	2	3	4	1+	2+	3+	4+
Arabic	28507.64	157.00597	3.793780	26.956297	34.83093329	9.301021	5.793780	28.956297	0.00000
Basque	10920.30	96.67550	15.446143	47.778695	19.65542112	27.808205	17.446143	49.778695	0.00000
Catalan	94789.90	273.00506	31.971886	116.927152	99.64206243	63.882555	33.971886	118.927152	0.00000
Chinese	160810.95	96.43779	12.464344	41.716466	31.79193509	23.706665	14.464344	43.716466	0.00000
Czech	129184.23	0.00000	1.794345	5.929212	7.82541669	1.999993	2.770618	7.929212	8.03038
English	110770.20	194.75858	14.792035	70.741753	47.17500214	32.510482	16.792035	72.741753	0.00000
Greek	18098.90	137.31024	8.323004	42.217897	26.98902499	18.403506	10.323004	44.217897	0.00000
Hungarian	35217.00	105.81917	8.487695	42.436947	0.02030926	19.097785	10.487695	44.436947	0.00000
Italian	23798.81	186.19586	16.572239	76.015730	43.57178949	34.835329	18.572239	78.015730	0.00000
Turkish	26832.38	144.73132	15.466098	56.596159	60.69399936	30.907310	17.466098	58.596159	0.00000

Figure 6: Δ AIC Table as per generated by our R Scripts.

Language	1b	2a	2b	3a	3c	4a	1+b	1+d	2+a	2+b	2+d	3+a	3+c	3+d	4+a	4+d
Arabic	0.1970271	2.249545	0.1970271	3.908448	0.003275220	1.8755846	0.1970271	0	2.249545	0.1970271	0	3.908448	0.003275220	0.0041323141	1.8755846	0
Basque	0.3521535	1.660602	0.3521535	3.198025	0.015949427	1.2318294	0.3521535	0	1.660602	0.3521535	0	3.198025	0.015949427	0.0058428050	1.2318294	0
Catalan	0.2338509	2.208456	0.2338509	4.036062	0.005004056	1.9570060	0.2338509	0	2.208456	0.2338509	0	4.036062	0.005004056	0.0148349042	1.9570060	0
Chinese	0.3017556	1.787342	0.3017556	3.201049	0.012761871	1.6784053	0.3017556	0	1.787342	0.3017556	0	3.201049	0.012761871	0.0012500000	1.6784053	0
Czech	0.3858377	1.433973	0.3858377	3.556052	0.009936070	-3.9669514	0.3858377	0	1.433973	0.3858377	0	3.556052	0.009936070	0.0037810208	-3.9669514	0
English	0.2512794	2.197358	0.2512794	4.157572	0.005562192	1.7962037	0.2512794	0	2.197358	0.2512794	0	4.157572	0.005562192	0.0003287048	1.7962037	0
Greek	0.2492269	2.076017	0.2492269	3.833453	0.005985845	1.6757035	0.2492269	0	2.076017	0.2492269	0	3.833453	0.005985845	0.0017578125	1.6757035	0
Hungarian	0.3549654	1.914675	0.3549654	4.359953	0.009842526	0.6208072	0.3549654	0	1.914675	0.3549654	0	4.359953	0.009842526	0.0609890110	0.6208072	0
Italian	0.2563963	2.027865	0.2563963	3.879665	0.005714291	1.6038420	0.2563963	0	2.027865	0.2563963	0	3.879665	0.005714291	0.0061727161	1.6038420	0
Turkish	0.2491276	1.986725	0.2491276	3.396096	0.008397338	1.9441683	0.2491276	0	1.986725	0.2491276	0	3.396096	0.008397338	0.0004573205	1.9441683	0

Figure 7: Models' initial values table as per generated by our R Scripts.

2.2 Final Visualisation

2.2.1 Arabic

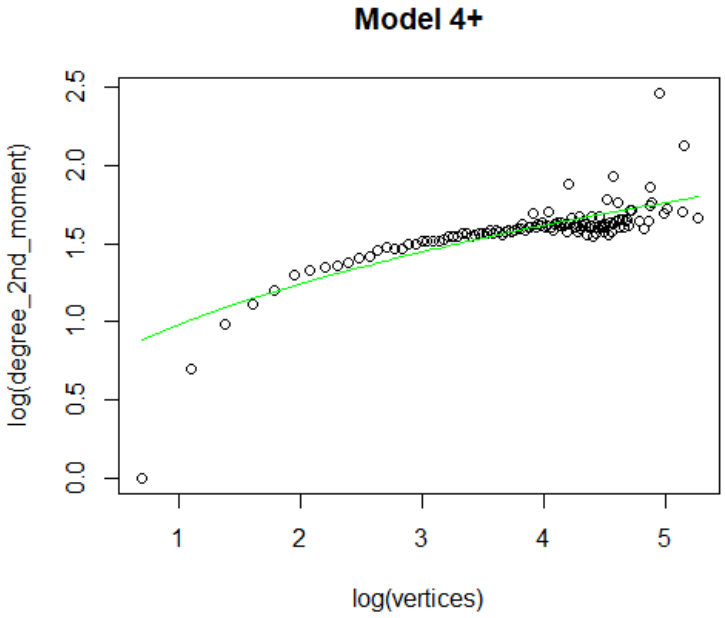

Figure 8: Arabic's best model as per our ΔAIC table is Model 4+.

2.2.2 Basque

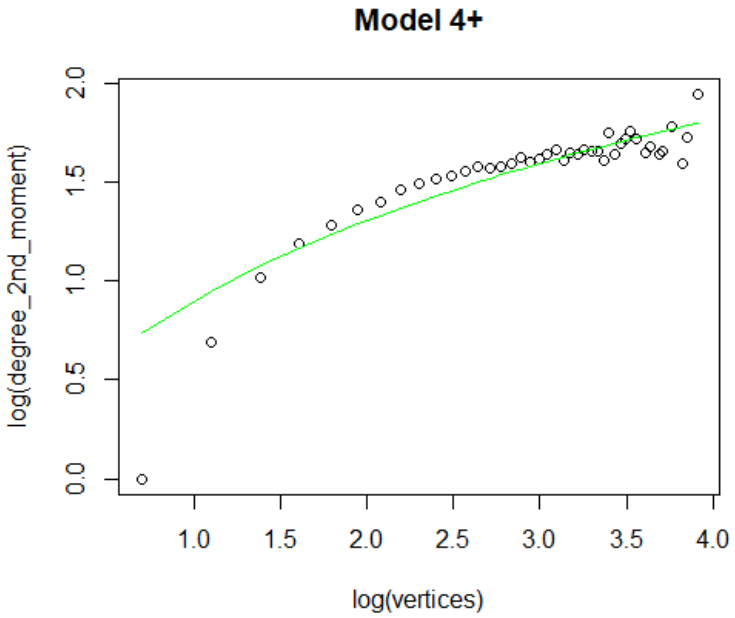

Figure 9: Basque's best model as per our ΔAIC table is Model 4+.

2.2.3 Catalan

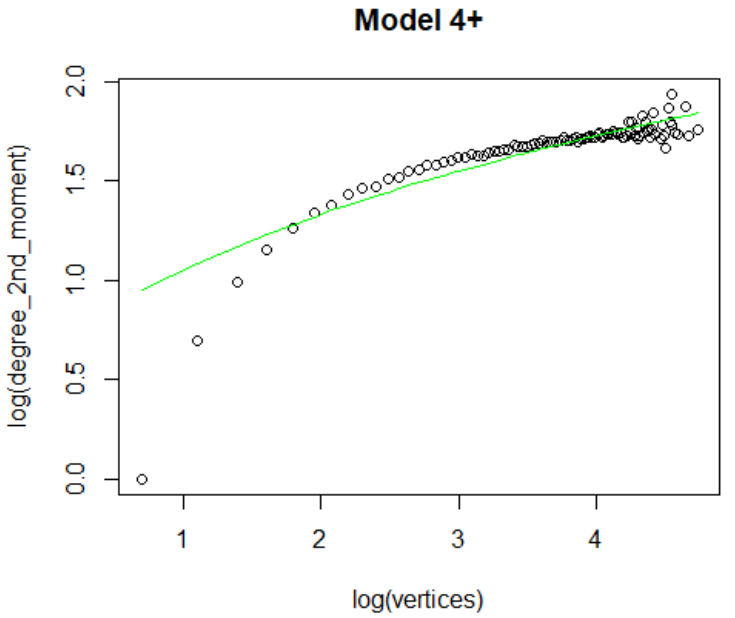

Figure 10: Catalan's best model as per our ΔAIC table is Model 4+.

2.2.4 Chinese

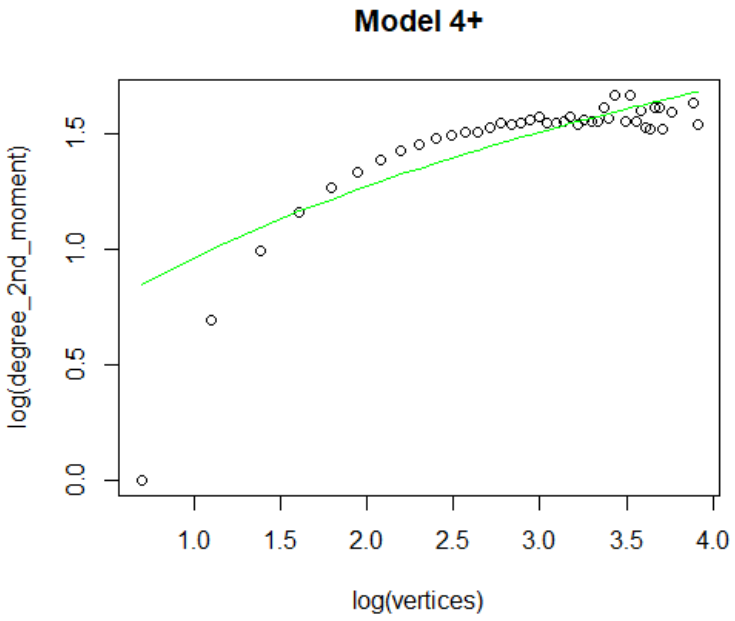

Figure 11: Chinese's best model as per our ΔAIC table is Model 4+.

2.2.5 Czech

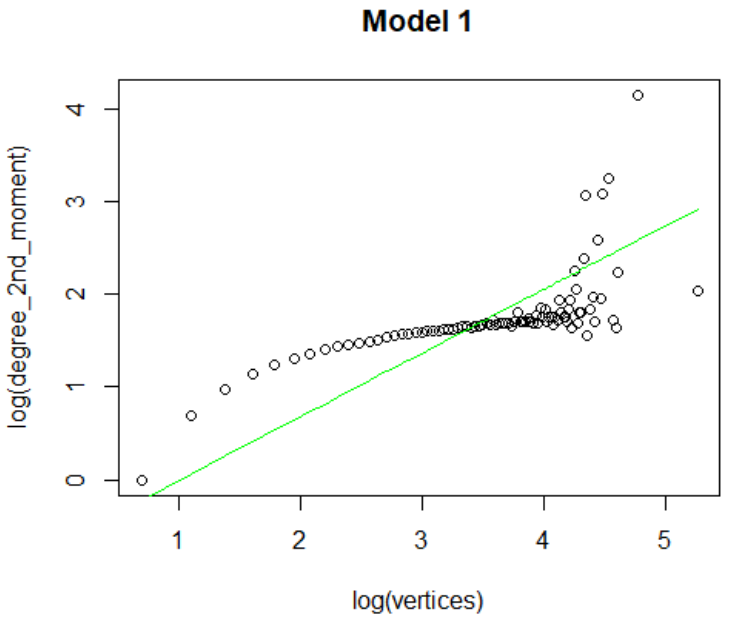

Figure 12: Czech's best model as per our ΔAIC table is Model 1.

2.2.6 English

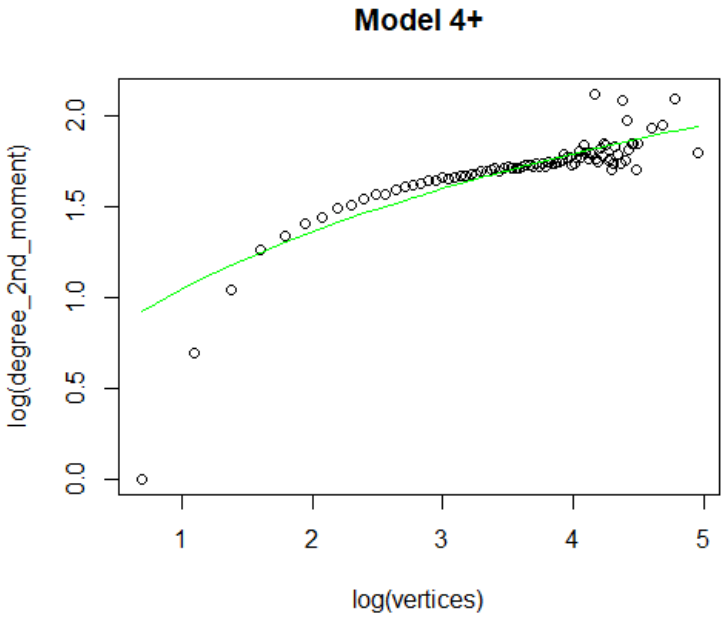

Figure 13: English's best model as per our ΔAIC table is Model 4+.

2.2.7 Greek

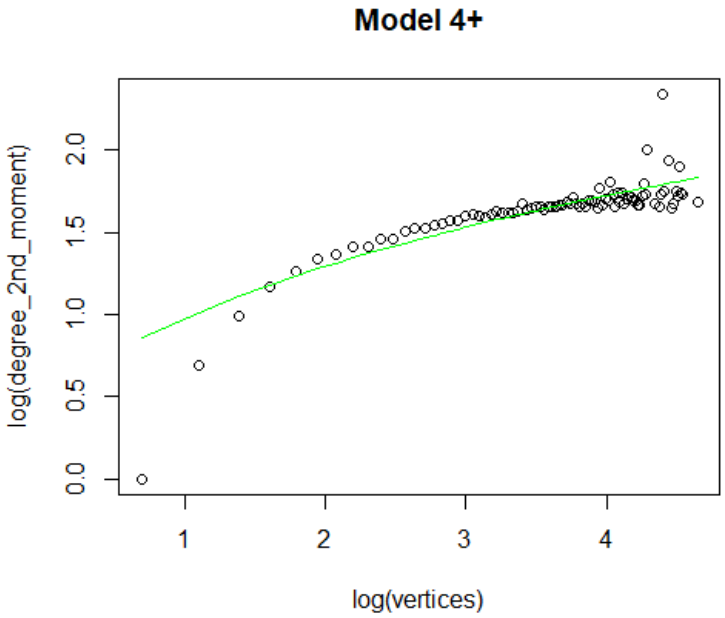

Figure 14: Greek's best model as per our ΔAIC table is Model 4+.

2.2.8 Hungarian

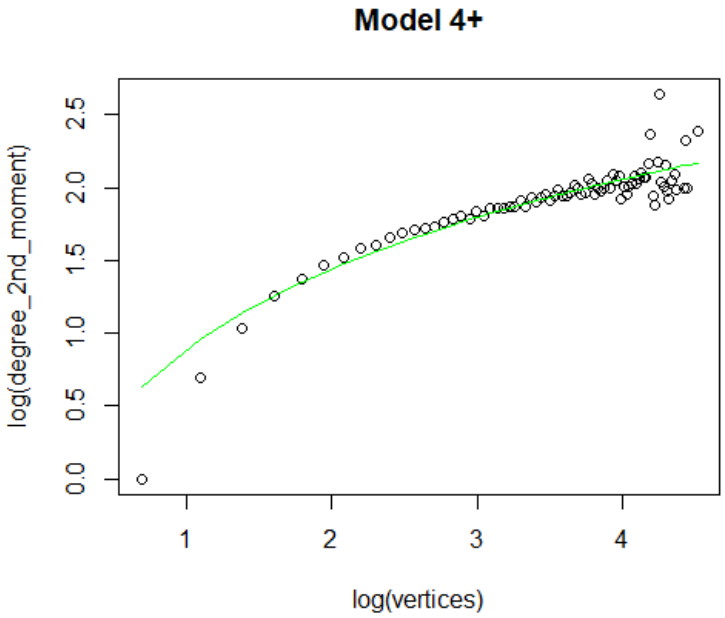

Figure 15: Hungarian's best model as per our ΔAIC table is Model 4+.

2.2.9 Italian

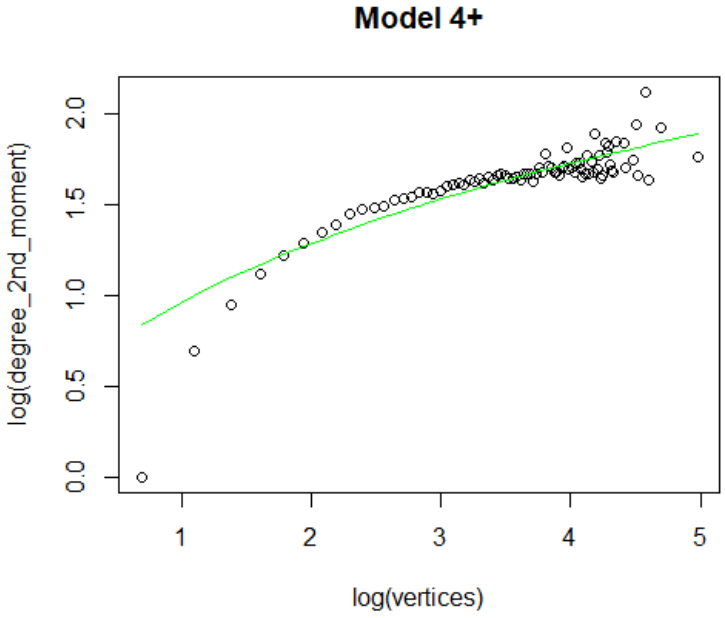

Figure 16: Italian's best model as per our ΔAIC table is Model 4+.

2.2.10 Turkish

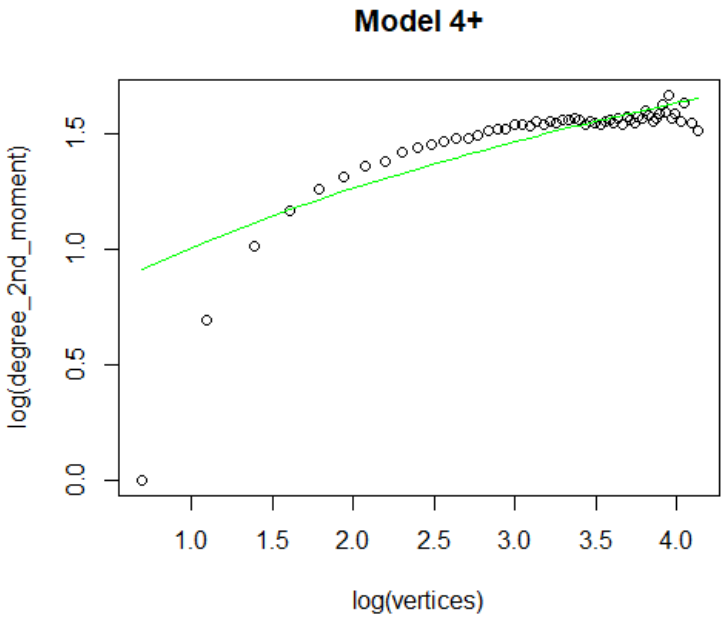

Figure 17: Turkish's best model as per our ΔAIC table is Model 4+.

3 Discussion

As noted by the table present in figure 6, the null hypothesis is always very far from the observed in the best model from all the languages.

In order to check for homoscedasticity for all the languages, we firstly calculate the variance in function of the number of vertices, then we calculate Fmax as per defined by the *Fmax / Hartley's Test*. Afterwards, we verify if $Fmax$ is close to 1 and, if it is, we can verify its homoscedasticity, otherwise, we proceed to obtain the number of participants per group (since the number of participants is, in reality, not the same in each group). To overcome this issue, we just take the mean of the number of participants per group. Other techniques could also be applied, however we opted for this one. As per Hartley's F_{max} Table actually, the table goes only to k = 12 and n = 60, and, on some cases we surpass this value quite largely, therefore, since on k = 12 and n = 60 tables indicate F_{max} should be ¡ 2.36, we merely check the F_{max} value in case of no homoscedasticity.

Regarding the debate over the model's best or not fit, we can agree that all languages have a very good fit with the exception of the Czech language, which was the only language where model 4+ was not the best fit, but, instead, the model 1. Furthermore, all languages are also very similar between each other, however, the Czech language has some very big outliers, which clearly destabilise the model selection.

4 Methods

Throughout the development of this lab's work we always strived to achieve a dynamic code, able to generate everything without any manual intervention with tables, graphs and values being calculated automatically. As such, when running the file $lab4.r$ the program generates the summary table and runs all the models for all the desired languages in one run.