```r
data <- fread(here("data", "summed_for_regression.csv"))
data[link_flair_text =="Computer Sci", link_flair_text := "Computer Science"]
print(head(data))
```

```
##       V1 jargon_proportion score num_comments link_flair_text  year month
##    <int>             <num> <int>        <num>          <char> <int> <int>
## 1:     0         0.1764706     1            0      Psychology  2017     6
## 2:     2         0.1714286    39            7      Psychology  2017     6
## 3:     4         0.0000000     2            0     Nanoscience  2017     6
## 4:     6         0.2857143     1            0       Chemistry  2017     6
## 5:     8         0.0000000     1            1        Medicine  2017     6
## 6:     9         0.1666667     1            2     Environment  2017     6
```

```r
print(str(data))
```

```
## Classes 'data.table' and 'data.frame':   198977 obs. of  7 variables:
##  $ V1               : int  0 2 4 6 8 9 11 12 14 15 ...
##  $ jargon_proportion: num  0.176 0.171 0 0.286 0 ...
##  $ score            : int  1 39 2 1 1 1 25 1 1 124 ...
##  $ num_comments     : num  0 7 0 0 1 2 5 0 0 19 ...
##  $ link_flair_text  : chr  "Psychology" "Psychology" "Nanoscience" "Chemistry" ...
##  $ year             : int  2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...
##  $ month            : int  6 6 6 6 6 6 6 6 6 6 ...
##  - attr(*, ".internal.selfref")=<externalptr>
##  - attr(*, "index")= int(0)
## NULL
```

```r
# count entries per category
print(data[, .N, by=link_flair_text])
```

```
##       link_flair_text     N
##                <char> <int>
##  1:        Psychology 15536
##  2:       Nanoscience  1943
##  3:         Chemistry  4476
##  4:          Medicine 16729
##  5:       Environment 15809
##  6:            Health 27802
##  7:           Physics 10238
##  8:  Computer Science  6178
##  9:      Anthropology  3726
## 10:         Astronomy 12381
## 11:    Social Science 11696
## 12:           Biology 20908
## 13:       Engineering  7104
## 14:      Paleontology  2812
## 15:    Animal Science  9558
## 16:      Neuroscience 10566
## 17:     Earth Science  6773
## 18:            Cancer  4656
## 19:       Mathematics  1060
## 20:           Geology  1997
## 21:      Epidemiology  4578
## 22:         Economics  1579
## 23:          Genetics   564
```

```
## 24: Materials Science   308
##       link_flair_text     N
```

```
model_jargon_only = lm(score ~ jargon_proportion, data=data)
model_year = lm(score ~ jargon_proportion + factor(year) + factor(month), data = data)
stargazer(model_jargon_only, model_year,
                     type="text",
                     omit = "factor",
                     column.labels = c("Jargon only", "Jargon, Year and Month")
                     )
```

```
##
## ================================================================================
##                                       Dependent variable:
##                     ---------------------------------------------------------
##                                             score
##                              Jargon only           Jargon, Year and Month
##                                  (1)                        (2)
## --------------------------------------------------------------------------------
## jargon_proportion             821.489***                820.942***
##                               (100.575)                  (100.471)
##
## Constant                      861.689***                145.285***
##                                (17.855)                  (52.965)
##
## --------------------------------------------------------------------------------
## Observations                   198,977                    198,977
## R2                              0.0003                     0.005
## Adjusted R2                     0.0003                     0.005
## Residual Std. Error  5,635.520 (df = 198975)     5,622.655 (df = 198958)
## F Statistic         66.715*** (df = 1; 198975) 55.311*** (df = 18; 198958)
## ================================================================================
## Note:                                            *p<0.1; **p<0.05; ***p<0.01
```

-> We can explain a bit more, when including factors for year months etc. Let's have a look whether that is significantly different:

```
anova(model_jargon_only, model_year)
```

```
## Analysis of Variance Table
##
## Model 1: score ~ jargon_proportion
## Model 2: score ~ jargon_proportion + factor(year) + factor(month)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1 198975 6.3193e+12
## 2 198958 6.2899e+12 17 2.9356e+10 54.623 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

-> Seems to be significantly better!

Let's look how the months influence our score? -> Seems like posts from December are the best ;)

```
stargazer::stargazer(model_jargon_only, model_year,
                     type="text",
                     omit="year"
                     )
```

```
##
## ============================================================================
##                                        Dependent variable:
##                     --------------------------------------------------------
##                                               score
##                                (1)                          (2)
## ----------------------------------------------------------------------------
## jargon_proportion           821.489***                   820.942***
##                             (100.575)                    (100.471)
##
## factor(month)2                                            -64.807
##                                                          (60.911)
##
## factor(month)3                                           -114.356*
##                                                          (59.153)
##
## factor(month)4                                            -44.952
##                                                          (61.121)
##
## factor(month)5                                            -84.353
##                                                          (60.063)
##
## factor(month)6                                           -100.308*
##                                                          (60.756)
##
## factor(month)7                                            -72.391
##                                                          (60.450)
##
## factor(month)8                                            -24.858
##                                                          (60.605)
##
## factor(month)9                                            20.284
##                                                          (64.760)
##
## factor(month)10                                           69.854
##                                                          (61.831)
##
## factor(month)11                                           71.156
##                                                          (61.785)
##
## factor(month)12                                         176.795***
##                                                          (62.039)
##
## Constant                    861.689***                   145.285***
##                              (17.855)                     (52.965)
##
## ----------------------------------------------------------------------------
## Observations                 198,977                      198,977
## R2                            0.0003                       0.005
## Adjusted R2                   0.0003                       0.005
## Residual Std. Error  5,635.520 (df = 198975)     5,622.655 (df = 198958)
## F Statistic         66.715*** (df = 1; 198975) 55.311*** (df = 18; 198958)
## ============================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Let's see how the year interacts with the jargon on the score:

```
data$year_factor = factor(data$year)
# Interaction terms in years and jargon
model_interaction_jargon_year = lm(score ~ jargon_proportion:year_factor + year_factor + factor(month) +
sanity_check = lm(score ~ jargon_proportion:year_factor + factor(year) + factor(month), data = data)
# THose should all be the same only moved by the intercept
stargazer::stargazer(model_interaction_jargon_year, sanity_check,
                     type="text",
                     omit=c("month", "jargon")
                     )
```

```
##
## ==========================================================================================
##                                                       Dependent variable:
##                                    -------------------------------------------------------
##                                                              score
##                                               (1)                          (2)
## ------------------------------------------------------------------------------------------
## year_factor2016                          211.782***
##                                           (59.468)
##
## year_factor2017                          613.556***
##                                           (63.894)
##
## year_factor2018                          888.337***
##                                           (63.936)
##
## year_factor2019                         1,236.686***
##                                           (63.902)
##
## year_factor2020                         1,136.631***
##                                           (61.326)
##
## year_factor2021                         1,070.569***
##                                           (62.243)
##
## year_factor2022                         1,033.864***
##                                           (62.570)
##
## factor(year)2017                                                    401.774***
##                                                                      (65.893)
##
## factor(year)2018                                                    676.555***
##                                                                      (65.290)
##
## factor(year)2019                                                   1,024.904***
##                                                                      (65.711)
##
## factor(year)2020                                                    924.849***
##                                                                      (62.642)
##
## factor(year)2021                                                    858.787***
##                                                                      (64.010)
##
```

```
## factor(year)2022                                                                822.083***
##                                                                                  (64.476)
##
## Constant                                                                        211.782***
##                                                                                  (59.468)
##
## ------------------------------------------------------------------------------------------
## Observations                                198,977                              198,977
## R2                                           0.033                                0.005
## Adjusted R2                                  0.033                                0.005
## Residual Std. Error (df = 198952)         5,622.617                            5,622.617
## F Statistic                        274.488*** (df = 25; 198952) 41.844*** (df = 24; 198952)
## ==========================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Okay seems like that works. Let's look at the interaction terms. I interpret this as how effective the jargon is in the different years.

```
stargazer(model_interaction_jargon_year,
                      type="text",
                      omit=1:18
                      )
```

```
##
## ================================================================
##                                      Dependent variable:
##                                  ---------------------------
##                                             score
## ----------------------------------------------------------------
## jargon_proportion:year_factor2016            282.271
##                                             (239.014)
##
## jargon_proportion:year_factor2017           768.854***
##                                             (276.666)
##
## jargon_proportion:year_factor2018           798.800***
##                                             (279.839)
##
## jargon_proportion:year_factor2019           990.915***
##                                             (291.505)
##
## jargon_proportion:year_factor2020          1,274.452***
##                                             (269.988)
##
## jargon_proportion:year_factor2021           979.151***
##                                             (264.185)
##
## jargon_proportion:year_factor2022           814.993***
##                                             (250.768)
##
## ----------------------------------------------------------------
## Observations                                 198,977
## R2                                            0.033
## Adjusted R2                                   0.033
## Residual Std. Error                  5,622.617 (df = 198952)
```

```
## F Statistic                         274.488*** (df = 25; 198952)
## ================================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

```
anova(model_interaction_jargon_year, model_year)
```

```
## Analysis of Variance Table
##
## Model 1: score ~ jargon_proportion:year_factor + year_factor + factor(month) -
##     1
## Model 2: score ~ jargon_proportion + factor(year) + factor(month)
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1 198952 6.2896e+12
## 2 198958 6.2899e+12 -6 -273175970 1.4402 0.1948
```

-> Also adding those interaction terms has a positive effect and better explains the variance.

When including interaction between jargon and years, we can improve model fit.

According to the anova test, this seems significant.

We could do something similar for the month, but I am not sure if it is worth it. Let's do it for the **categories**.

```
# add a count to unique values
data$fac_category = factor(data$link_flair_text)
unique(data$fac_category)
```

```
##  [1] Psychology        Nanoscience       Chemistry         Medicine
##  [5] Environment       Health            Physics           Computer Science
##  [9] Anthropology      Astronomy         Social Science    Biology
## [13] Engineering       Paleontology      Animal Science    Neuroscience
## [17] Earth Science     Cancer            Mathematics       Geology
## [21] Epidemiology      Economics         Genetics          Materials Science
## 24 Levels: Animal Science Anthropology Astronomy Biology Cancer ... Social Science
```

```
# I'm not sure whether I can do it in the following way, but it would be easier to interpret:
# It gives basically the same results but shifted. For some reason the F values are different though.
model_interaction_jargon_category = lm(score ~ jargon_proportion*fac_category - jargon_proportion - 1 +
stargazer(model_interaction_jargon_category,
          type="text",
          omit="factor")
```

```
##
## ================================================================================
##                                                    Dependent variable:
##                                                ---------------------------
##                                                           score
## --------------------------------------------------------------------------------
## fac_categoryAnimal Science                                52.947
##                                                          (90.855)
##
## fac_categoryAnthropology                                296.797**
##                                                         (136.101)
##
## fac_categoryAstronomy                                  -432.483***
##                                                          (81.234)
##
## fac_categoryBiology                                       75.213
```

```
##                                                  (74.048)
##
## fac_categoryCancer                              279.595**
##                                                 (130.461)
##
## fac_categoryChemistry                           -104.689
##                                                 (121.081)
##
## fac_categoryComputer Science                   -425.935***
##                                                 (102.222)
##
## fac_categoryEarth Science                      -354.235***
##                                                  (99.144)
##
## fac_categoryEconomics                           808.215***
##                                                 (213.951)
##
## fac_categoryEngineering                         -109.436
##                                                 (103.337)
##
## fac_categoryEnvironment                         382.783***
##                                                  (83.105)
##
## fac_categoryEpidemiology                        726.858***
##                                                 (119.795)
##
## fac_categoryGenetics                             -11.416
##                                                 (330.309)
##
## fac_categoryGeology                             -262.413
##                                                 (180.204)
##
## fac_categoryHealth                              303.691***
##                                                  (71.090)
##
## fac_categoryMaterials Science                    398.039
##                                                 (426.655)
##
## fac_categoryMathematics                         -513.605**
##                                                 (219.191)
##
## fac_categoryMedicine                            218.588***
##                                                  (84.515)
##
## fac_categoryNanoscience                          -62.506
##                                                 (166.766)
##
## fac_categoryNeuroscience                        578.022***
##                                                  (95.922)
##
## fac_categoryPaleontology                          82.949
##                                                 (155.870)
##
## fac_categoryPhysics                            -332.271***
```

```
##                                                    (91.110)
##
## fac_categoryPsychology                           1,525.441***
##                                                    (89.532)
##
## fac_categorySocial Science                        943.623***
##                                                    (92.698)
##
## jargon_proportion:fac_categoryAnimal Science       -31.193
##                                                   (626.005)
##
## jargon_proportion:fac_categoryAnthropology         787.783
##                                                  (1,137.373)
##
## jargon_proportion:fac_categoryAstronomy            601.048
##                                                   (466.096)
##
## jargon_proportion:fac_categoryBiology             -395.941
##                                                   (291.702)
##
## jargon_proportion:fac_categoryCancer              -742.321
##                                                   (591.426)
##
## jargon_proportion:fac_categoryChemistry           1,217.963*
##                                                   (665.338)
##
## jargon_proportion:fac_categoryComputer Science     189.624
##                                                   (543.863)
##
## jargon_proportion:fac_categoryEarth Science        481.462
##                                                   (595.208)
##
## jargon_proportion:fac_categoryEconomics          6,916.429***
##                                                  (1,194.317)
##
## jargon_proportion:fac_categoryEngineering          274.181
##                                                   (449.956)
##
## jargon_proportion:fac_categoryEnvironment          559.484
##                                                   (391.433)
##
## jargon_proportion:fac_categoryEpidemiology         -60.614
##                                                  (1,100.985)
##
## jargon_proportion:fac_categoryGenetics             -70.185
##                                                  (2,038.196)
##
## jargon_proportion:fac_categoryGeology              386.407
##                                                  (1,155.254)
##
## jargon_proportion:fac_categoryHealth             1,225.557***
##                                                   (239.392)
##
## jargon_proportion:fac_categoryMaterials Science    -621.884
```

```
##                                                        (2,338.220)
##
## jargon_proportion:fac_categoryMathematics                 96.319
##                                                        (1,496.166)
##
## jargon_proportion:fac_categoryMedicine                  691.764**
##                                                         (300.689)
##
## jargon_proportion:fac_categoryNanoscience               -281.972
##                                                        (1,248.226)
##
## jargon_proportion:fac_categoryNeuroscience            -1,035.720**
##                                                         (513.231)
##
## jargon_proportion:fac_categoryPaleontology              -406.465
##                                                        (1,316.523)
##
## jargon_proportion:fac_categoryPhysics                     17.861
##                                                         (403.208)
##
## jargon_proportion:fac_categoryPsychology              1,544.273***
##                                                         (436.207)
##
## jargon_proportion:fac_categorySocial Science          2,018.564***
##                                                         (477.644)
##
## ------------------------------------------------------------------------------
## Observations                                              198,977
## R2                                                          0.044
## Adjusted R2                                                 0.043
## Residual Std. Error                           5,592.869 (df = 198912)
## F Statistic                                 139.961*** (df = 65; 198912)
## ==============================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```