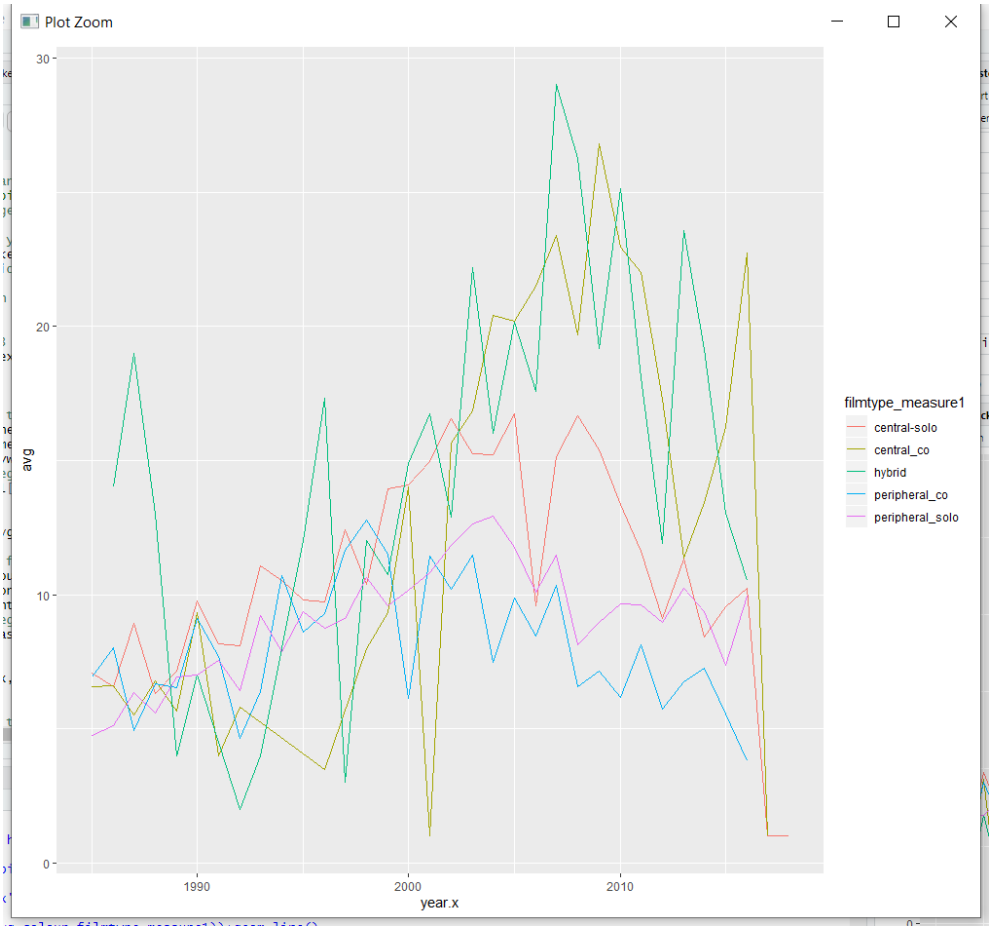
Q1.

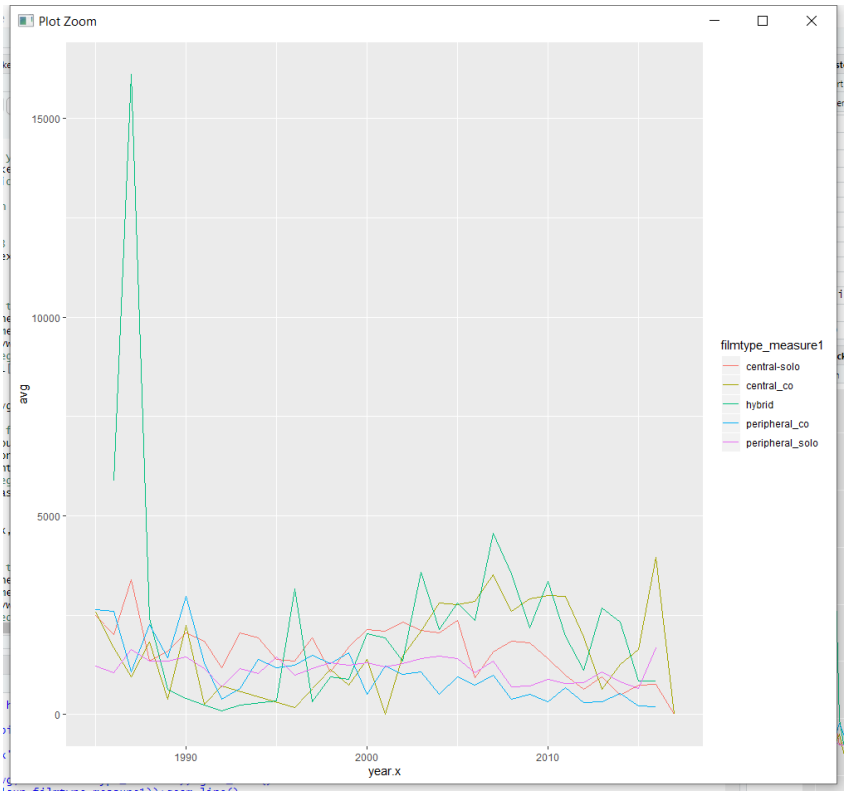
A.

Classify each film by the type of collaboration that it represents. There should be five types for each measure of generalism:
i. Peripheral solo productions: films made by a single specialist
ii. Central solo productions: films made by a single generalist
iii. Central co-productions: films made by a group of multiple generalists
iv. Peripheral co-productions: films made by a group of multiple specialists
v. Hybrid co-productions: films made by a group of generalists and specialists
For each measure of generalism, a figure that illustrates the number of new keywords and new combinations of existing keywords that are introduced per type of film over the course of the data. On the *x*-axis should be years, and on the *y*-axis should be the count of new keywords or new combinations.

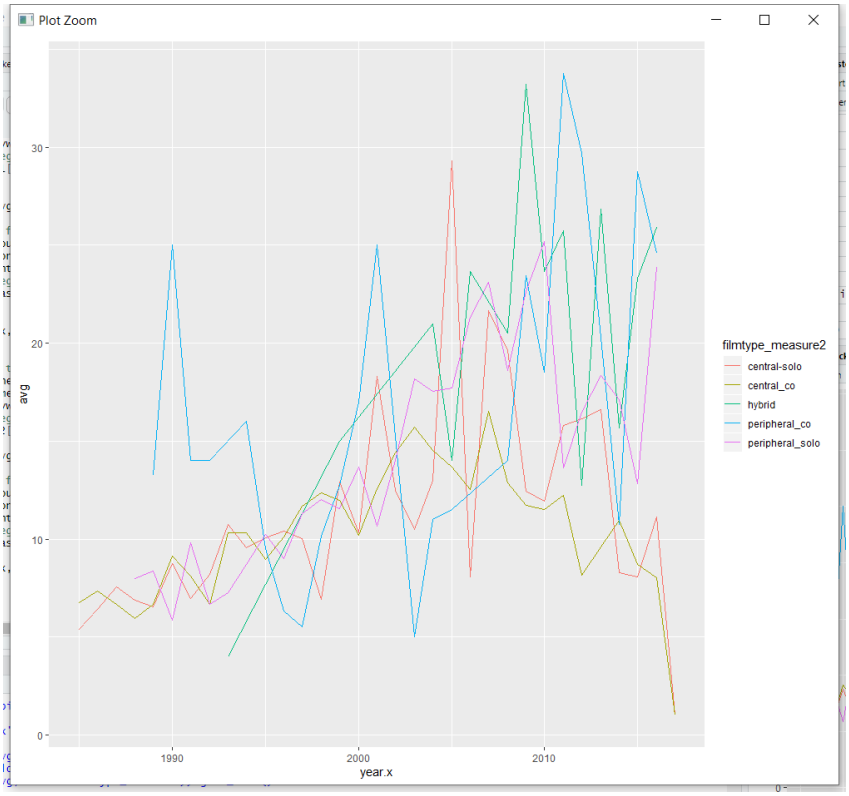
Measure1 + only single new keyword



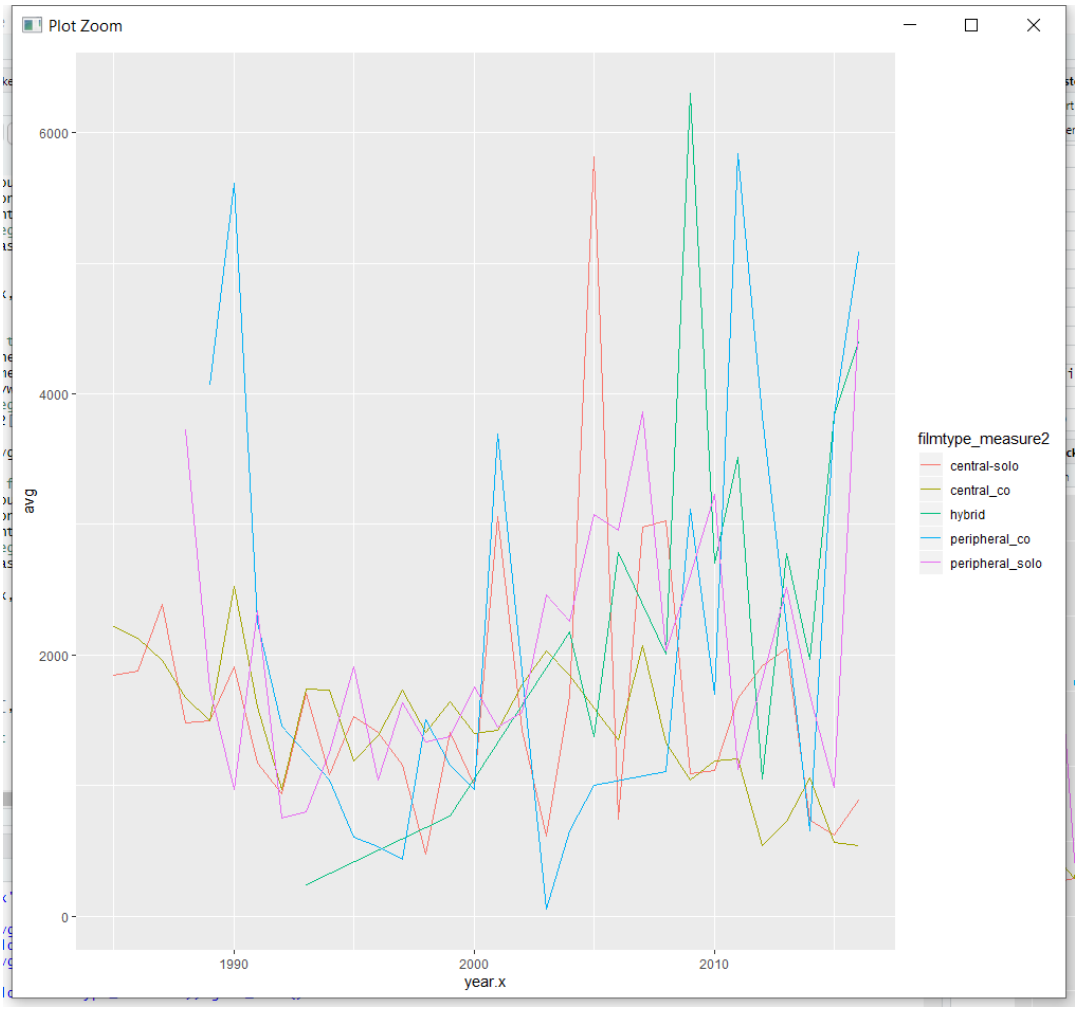
Measure 1 + new key words combination



Measure 2 + new key words



Measure 2 + new key word combination



B.

What kinds of collaborations seem to result in the most new keywords and new combinations of existing keywords? Comparing the two measures of generalism, are collaborations between large or small companies (measure1)or core and peripheral companies(measure2) more effective for creative innovation?

1.single key word + measure 1

```
Call:
glm.nb(formula = N ~ central_co_prod + peripheral_co_prod + hybrid_co +
    V1 + V2 + rev_prod + numberofyear + issub + factor(year),
    data = reg2.1, weights = offset(totalfilm), init.theta = 1.563992775,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-72.578  -22.208   -7.215    7.877  102.157

Coefficients:
                     Estimate Std. Error  z value Pr(>|z|)
(Intercept)         4.938e+00  7.402e-03  667.108   <2e-16 ***
central_co_prod     3.189e-01  1.245e-03  256.035   <2e-16 ***
peripheral_co_prod -1.952e-01  1.575e-03 -123.910   <2e-16 ***
hybrid_co           3.059e-01  1.594e-03  191.924   <2e-16 ***
V1                 -1.299e-01  2.500e-03  -51.945   <2e-16 ***
V2                 -1.391e-02  4.408e-03   -3.155   0.0016 **
rev_prod            1.492e-10  3.771e-13  395.556   <2e-16 ***
numberofyear        1.411e-02  1.130e-04  124.846   <2e-16 ***
issub               4.593e-01  2.359e-03  194.687   <2e-16 ***
factor(year)1986   -3.779e-01  9.585e-03  -39.427   <2e-16 ***
factor(year)1987   -2.393e-01  9.427e-03  -25.384   <2e-16 ***
factor(year)1988   -4.255e-01  9.148e-03  -46.517   <2e-16 ***
```

2. single key word + measure 2

```
Call:
glm.nb(formula = N ~ central_co_prod + peripheral_co_prod + hybrid_co +
    V1 + V2 + rev_prod + numberofyear + issub + factor(year),
    data = reg2.2, weights = offset(totalfilm), na.action = "na.omit",
    init.theta = 1.610354128, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-67.222  -21.646   -6.433    7.567  127.892

Coefficients:
                     Estimate Std. Error  z value Pr(>|z|)
(Intercept)         4.512e+00  7.821e-03  576.874  < 2e-16 ***
central_co_prod     1.451e-01  8.131e-04  178.422  < 2e-16 ***
peripheral_co_prod  1.553e-01  2.742e-03   56.648  < 2e-16 ***
hybrid_co           4.100e-01  2.887e-03  142.028  < 2e-16 ***
V1                 -2.121e-01  2.680e-03  -79.115  < 2e-16 ***
V2                  1.091e-01  4.580e-03   23.823  < 2e-16 ***
rev_prod            1.536e-10  3.876e-13  396.264  < 2e-16 ***
numberofyear        1.506e-02  1.167e-04  129.072  < 2e-16 ***
issub               4.267e-01  2.367e-03  180.249  < 2e-16 ***
factor(year)1986    5.633e-02  1.027e-02    5.482  4.2e-08 ***
factor(year)1987    2.083e-01  1.041e-02   20.012  < 2e-16 ***
```

3.key word combination + measure1

```
glm.nb(formula = count ~ central_co_prod + peripheral_co_prod +
    hybrid_co + V1 + V2 + rev_prod + numberofyear + issub + factor(year),
    data = reg2.3, weights = offset(totalfilm), init.theta = 0.4172884553,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-78.315  -25.980   -9.670    2.925  115.930

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.084e+01  1.421e-02  762.77   <2e-16 ***
central_co_prod     5.443e-01  2.252e-03  241.70   <2e-16 ***
peripheral_co_prod -4.731e-01  2.527e-03 -187.20   <2e-16 ***
hybrid_co           4.890e-01  2.759e-03  177.25   <2e-16 ***
V1                  5.272e-01  4.022e-03  131.08   <2e-16 ***
V2                 -2.363e-01  7.270e-03  -32.50   <2e-16 ***
rev_prod            3.034e-10  6.849e-13  442.96   <2e-16 ***
numberofyear        1.188e-01  1.848e-04  642.95   <2e-16 ***
issub               7.397e-01  4.222e-03  175.20   <2e-16 ***
factor(year)1986   -9.976e-01  1.842e-02  -54.16   <2e-16 ***
factor(year)1987   -1.025e+00  1.811e-02  -56.58   <2e-16 ***
factor(year)1988   -4.797e-01  1.760e-02  -27.26   <2e-16 ***
```

4. key word combination + measure2

```
Call:
glm.nb(formula = count ~ central_co_prod + peripheral_co_prod +
    hybrid_co + V1 + V2 + rev_prod + numberofyear + issub + factor(year),
    data = reg2.4, weights = offset(totalfilm), init.theta = 0.4726355868,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-84.984  -24.359   -8.612    4.082  104.796

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.025e+01  1.432e-02  715.72   <2e-16 ***
central_co_prod     3.808e-02  1.398e-03   27.23   <2e-16 ***
peripheral_co_prod  3.008e-01  4.663e-03   64.51   <2e-16 ***
hybrid_co           6.171e-01  4.888e-03  126.24   <2e-16 ***
V1                  3.445e-01  4.149e-03   83.03   <2e-16 ***
V2                 -2.933e-01  7.294e-03  -40.22   <2e-16 ***
rev_prod            3.277e-10  6.715e-13  488.02   <2e-16 ***
numberofyear        1.020e-01  1.859e-04  548.58   <2e-16 ***
issub               7.085e-01  4.062e-03  174.41   <2e-16 ***
factor(year)1986   -5.594e-01  1.882e-02  -29.72   <2e-16 ***
factor(year)1987   -3.859e-01  1.907e-02  -20.24   <2e-16 ***
```

Answer:

1. Regarding new single key words, all three collaborations are highly correlated with the number of new keywords generated as the p-value of each of them is very small (p<2e-16) in the summary of regression. And the result holds for using either measure 1 or measure 2.
    a. When using measure 1, central co-production and hybrid co-production have positive coefficient (0.3189 and 0.3059) but peripheral co-production has negative one (-0.1952). Between central and hybrid, central co-production has higher coefficient but the difference is very small. When using measure 2, all three types have positive coefficients (0.1451, 0.1553, 0.41) and hybrid co-production has the highest one, which is two times larger than the other two types' coefficient. So that, when considering single key words, hybrid co-production seems to result into most new keywords. Hybrid co-production is among a group of generalists and specialists. It's reasonable that when team members are from different background and the number of people is very large it's much more possible to generate innovative ideas.

Regarding new keywords combination, all three types of collaborations are highly correlated with the number of new keywords combinations as their p-value is very small ($p<2e-16$). The result holds true for using either measure 1 or measure 2.

a. When using measure 1, central co-production and hybrid have positive coefficients (0.5443 and 0.4890) but peripheral co-production has negative one (-0.4731). The difference between central and hybrid is very small. When using measure, all three types have positive coefficients (0.038, 0.3008, 0.617), among which hybrid co-production has the highest value. So that when considering keywords combinations, it's also hybrid co-production that results into most new stuff, which is words combination here. The reason behind it is the same as in the case of counting single keyword.

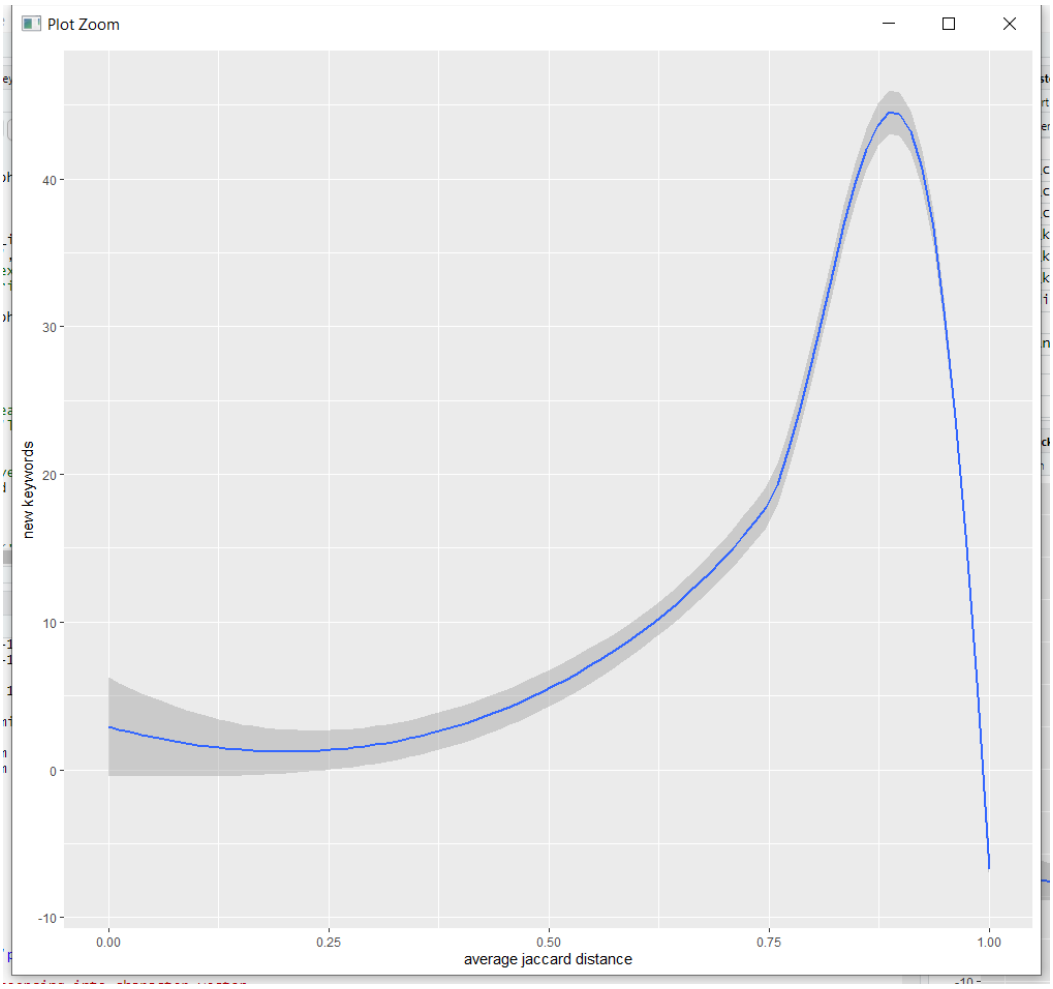
2.

Coefficient	Central co-production	Peripheral co-production	hybrid
Single keywords Measure1	0.3189	-0.1952	0.3059
Single keywords Measure2	0.1451	0.1553	0.4100
Combination Measure1	0.5443	-0.4731	0.4890
Combination Measure2	0.03808	0.3008	0.6171

For central co-production, the coefficient is larger when using measure 1, which estimates collaborations between large and small companies. For hybrid co-production, the coefficient is larger when using measure 2, which estimate collaborations between core and peripheral companies. As hybrid co-production is most effective in generating innovation as discussed above. I would say Collaborations between core and peripheral companies are more effective for creative innovations.

Q2

What does the pattern suggest about what kinds of collaborative partnerships might result in more creative innovation? Does this help to explain the results from Question 1?



Answer:

The pattern of the trend line looks like an arc, which reaches the highest value at the center-right place. It means that producers can not generate much innovative creations when their jaccard distance is either too small or too big, in other words, they are too similar or too dissimilar. The largest innovation is achieved when producers are not so similar but know each other. Among the three types of partnerships, hybrid co-production is the most proper one to maintain this moderate-distance relationship. When the partnership is hybrid co-production in which a group of generalists and specialists cooperate together, it's very possible to have some similarity between generalist and specialist as generalist may make films in many topics. But the similarity will not be too big as specialist only focuses on a few areas. So the hybrid co-production perfectly meet the requirements for moderate-distance partnership. That's also why in

Question 1 the number of hybrid co-production a producer has is most positively related to the number of new keywords or keywords combination the producer introduces in a year.

Q3.

Next, let's analyze whether collaborations influence a production company's financial returns. Estimate a regression predicting producers' standardized return using the core-periphery classification to define generalists and specialists.

What do the results suggest about financial outcomes for collaborations?

```
Call:
lm(formula = total_norm ~ central_co_prod + peripheral_co_prod +
    hybrid_co + V1 + V2 + rev_prod + numberofyear + issub + factor(year),
    data = reg3.2)

Residuals:
    Min      1Q  Median      3Q     Max
-41.704  -2.189  -0.126   3.435  39.628

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -3.220e+00  1.539e+00  -2.093 0.036465 *
central_co_prod    -9.128e-01  2.070e-01  -4.409 1.08e-05 ***
peripheral_co_prod  2.128e+00  8.250e-01   2.580 0.009937 **
hybrid_co          -7.207e-01  9.769e-01  -0.738 0.460710
V1                  1.669e+00  7.467e-01   2.235 0.025480 *
V2                  7.822e-01  1.278e+00   0.612 0.540530
rev_prod            1.233e-08  9.820e-11 125.558  < 2e-16 ***
numberofyear       -4.431e-01  3.439e-02 -12.886  < 2e-16 ***
issub              -7.473e+00  6.227e-01 -12.002  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.754 on 2853 degrees of freedom
Multiple R-squared:  0.8873,    Adjusted R-squared:  0.8858
F-statistic: 575.9 on 39 and 2853 DF,  p-value: < 2.2e-16
```

Answer:

Look at the regression result, we can see the adjusted r-squared is 0.8858 and p-value of the whole model is lower than 2.2e-16, so that this model could well estimate the standardized return of producers. Then look at each predictor. Variable number of central co-production and peripheral co-production both have very small p-value (1.08e-05 and 0.009937 respectively). But the number of hybrid co-production is not very related as the p-value is big (0.46071). The coefficient for central co-production and hybrid co-production are both negative and the coefficient for peripheral co-production is positive, which means a group of generalists or a mix of both generalists and specialists would hurt the financial return on the film but a group of specialists would increase the financial return on the film. As we are predicting financial return using the same year value of predictors, it seems like most collaborations would hurt the company in the short run. The reason is quite realistic. Because at the beginning, producers need to take time to integrate the background and experience of each other, which may take some time and pain. That's why both central co-production and hybrid co-production have negative coefficients. Regarding peripheral co-production, it's different. Specialization can always create high-quality creations and when specialists work together, they usually specialize in the same area. So a group of specialists working on a same film

are more likely to generate good works in a short period of time. On the contrast, generalist can do many things but are not specialized on each of them. Based on our definition, generalists are companies producing many films per year so that they can not spend too much time on each of them. So generalists working together are hard to produce high-quality works, which is another reason that central co-production has negative coefficient.

Q4

A.

Estimate a regression predicting the count of new keywords introduced in a producer's *solo* produced films in a year.

Does creative innovation gained through collaborations make a producer's solo-produced films more innovative? What does this suggest?

```
Call:
glm.nb(formula = count_new_solo ~ cum_hybrid + V1 + V2 + rev_prod +
    numberofyear + issub + factor(year), data = reg4.1, weights = offset(totalfilm),
    init.theta = 0.7993257109, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-77.696  -23.514   -9.095    5.852  128.680

Coefficients:
                   Estimate Std. Error  z value Pr(>|z|)
(Intercept)       7.371e+00  7.000e-03 1053.082   <2e-16 ***
cum_hybrid        8.056e-02  1.090e-04  739.027   <2e-16 ***
V1               -1.860e-01  2.799e-03  -66.464   <2e-16 ***
V2               -3.456e-01  5.321e-03  -64.947   <2e-16 ***
rev_prod          4.489e-10  3.426e-13 1310.213   <2e-16 ***
numberofyear     -2.329e-05  8.430e-08 -276.228   <2e-16 ***
issub             1.066e+00  1.906e-03  559.455   <2e-16 ***
factor(year)1986 -3.461e-01  9.236e-03  -37.471   <2e-16 ***
```

Answer:

Yes, creative innovation does make a producer's solo-produced films more innovative as the coefficient for number of cumulative new keywords made in hybrid films is positive. So that even many collaborations may be risky in the short run as discussed question 3, many producers still resort to collaboration as it can generate innovation in the long run.

B.

Accounting for a producer's engaging in collaborations, does introducing new keywords result in higher box office returns?

Does this result help explain why producers might engage in collaborations, even though they can be financially risky

Measure1:

```
Call:
lm(formula = total_norm ~ N + central_co_prod + peripheral_co_prod + 
    hybrid_co + V1 + V2 + rev_prod + numberofyear + issub + factor(year), 
    data = reg4.21)

Residuals:
    Min      1Q  Median      3Q     Max 
-59.910  -2.898  -0.310   3.201  65.143 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         3.082e+00  1.936e+00   1.592  0.11146    
N                   1.752e-08  1.238e-10 141.510  < 2e-16 ***
central_co_prod     1.481e+00  4.660e-01   3.179  0.00149 ** 
peripheral_co_prod -2.025e+00  4.450e-01  -4.550 5.55e-06 ***
hybrid_co          -8.958e-01  5.948e-01  -1.506  0.13217    
V1                  2.277e+00  8.501e-01   2.678  0.00743 ** 
V2                  5.483e-01  1.494e+00   0.367  0.71367    
rev_prod           -3.887e-02  4.528e-03  -8.585  < 2e-16 ***
numberofyear       -5.341e-01  4.111e-02 -12.993  < 2e-16 ***
issub              -1.222e+01  7.595e-01 -16.086  < 2e-16 ***
factor(year)1986   -3.171e+00  2.313e+00  -1.371  0.17043    
factor(year)1987   -1.975e+00  2.383e+00  -0.829  0.40724    

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.13 on 3285 degrees of freedom
Multiple R-squared:  0.902,     Adjusted R-squared:  0.9009 
F-statistic: 756.3 on 40 and 3285 DF,  p-value: < 2.2e-16
```

Measure2:

```
Call:
lm(formula = total_norm ~ N + central_co_prod + peripheral_co_prod + 
    hybrid_co + V1 + V2 + rev_prod + numberofyear + issub + factor(year), 
    data = reg4.22)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.521  -3.083  -0.506   3.518  68.372 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         2.098e+00  2.202e+00   0.953 0.340789    
N                   1.783e-08  1.353e-10 131.734  < 2e-16 ***
central_co_prod    -8.796e-01  2.777e-01  -3.167 0.001556 ** 
peripheral_co_prod  3.907e+00  1.081e+00   3.616 0.000305 ***
hybrid_co          -2.084e+00  1.279e+00  -1.629 0.103349    
V1                  1.910e+00  9.780e-01   1.953 0.050888 .  
V2                  7.995e-01  1.673e+00   0.478 0.632827    
rev_prod           -3.410e-02  5.053e-03  -6.747 1.82e-11 ***
numberofyear       -5.490e-01  4.521e-02 -12.143  < 2e-16 ***
issub              -1.259e+01  8.179e-01 -15.387  < 2e-16 ***
factor(year)1986   -1.465e+00  2.643e+00  -0.554 0.579520    
factor(year)1987   -1.098e+00  2.839e+00  -0.387 0.698979    

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.77 on 2852 degrees of freedom
Multiple R-squared:  0.9045,    Adjusted R-squared:  0.9031 
F-statistic:   675 on 40 and 2852 DF,  p-value: < 2.2e-16
```

Answer:

Yes, it does explain. According to the regression result of using measure 1, we can see number of new keywords is highly and positively related to the financial return of the producer (p<2e-15, coefficient=1.752e-08). When considering measure 2, the relationship is also strong and positive (p-value<2e-16, coefficient=1.783e-08). That's why even though collaboration generates short-term financial loss, many producers still do collaboration. Because introducing new keywords generated by collaboration can positively influence the financial return of the company in the long run.

Extra Credits:

Does engaging in more hybrid collaborations seem to help with hiring more innovative creative talent?

```
Call:
glm.nb(formula = cum_prod ~ central_co_prod + peripheral_co_prod +
    hybrid_co, data = pred_inno, weights = offset(totalfilm),
    init.theta = 0.298065732, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-75.281  -28.074  -10.787    1.395  107.899

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       7.115311   0.001211  5877.3   <2e-16 ***
central_co_prod   1.124027   0.002310   486.6   <2e-16 ***
peripheral_co_prod -0.777783   0.002199  -353.8   <2e-16 ***
hybrid_co         1.042779   0.002947   353.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.2981) family taken to be 1)

    Null deviance: 5418019  on 5773  degrees of freedom
Residual deviance: 4655839  on 5770  degrees of freedom
AIC: 54922553

Number of Fisher Scoring iterations: 1


              Theta:  0.298066
          Std. Err.:  0.000193

 2 x log-likelihood:  -54922543.199000
```

Answer:

Yes, engaging in more hybrid collaborations seem to help with hiring more innovative creative talent as the coefficient of hybrid co-production is positive and the p-value is pretty small. Hybrid co-production can let generalists meet specialists, which provides a good opportunity for talents to flow into other companies.