BUS 41201 Homework 7 Assignment

Group 11

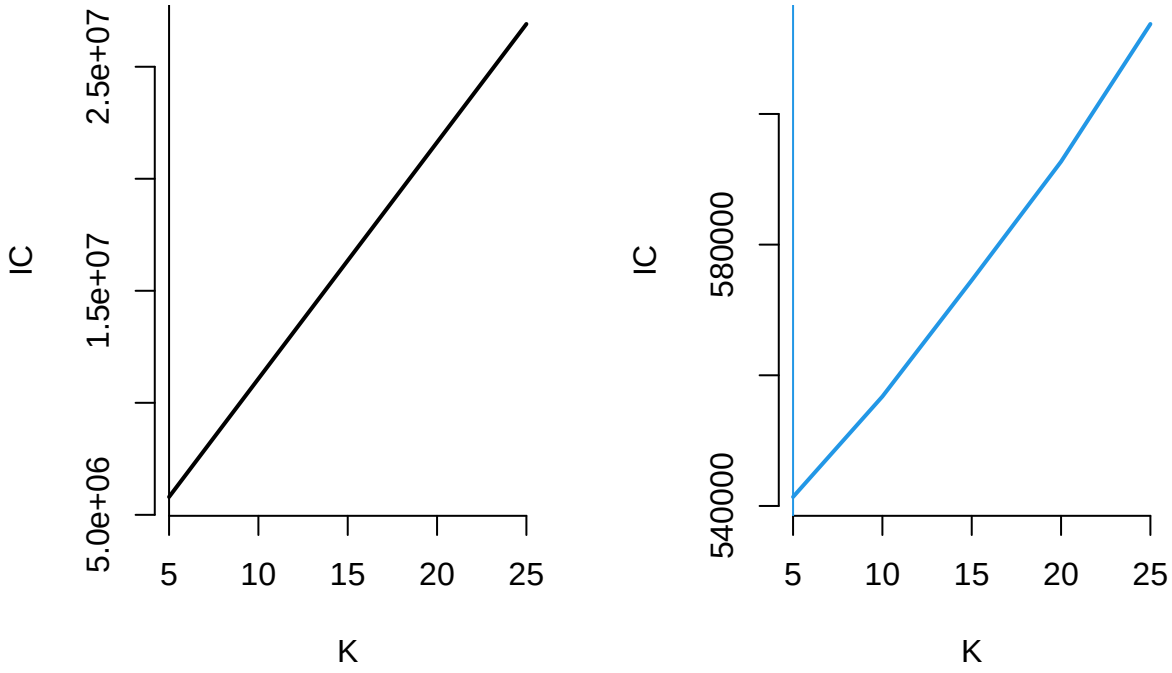
2024-05-05

1. Fit K-means to speech text for K in 5,10,15,20,25. Use BIC to choose the K and interpret the selected model.

```
data(congress109)
set.seed(123)
#  as we discussed in class, you can choose a variety of scales
#  upon which to fit k-means; here I just used standardized freq
fs <- scale(as.matrix( congress109Counts/rowSums(congress109Counts) ))

## follow wine code to fit for a set of k's
## notice the only difference here is that I've replaced 1:200 with 5*(1:5)
## the question asked for a smaller set of candidate models than we had for wine.
kfit <- lapply(5*(1:5), function(k) kmeans(fs,k))
```

```
setwd("C:/Users/user/Desktop/Big Data/HW/week7")
source("kIC.R")

kaicc <- sapply(kfit,kIC)
kbic <- sapply(kfit,kIC,"B")
```

```
## plot 'em:
par(mfrow=c(1,2))
plot(5*(1:5), kaicc, xlab="K", ylab="IC",
    bty="n", type="l", lwd=2)
abline(v=which.min(kaicc)*5)
plot(5*(1:5), kbic, xlab="K", ylab="IC",
    bty="n", type="l", lwd=2, col=4)
abline(v=which.min(kbic)*5,col=4)
```




```
k_values <- c(5,10,15,20,25)
optimal_k_aicc <- k_values[which.min(kaicc)]
optimal_k_bic <- k_values[which.min(kbic)]
optimal_k_bic
```

```
## [1] 5
```

```
kmfs <- kfit[[which(k_values==optimal_k_bic)]]
## interpretation: we can see the words with cluster centers
## highest above zero (these are in units of standard deviation of f)
print(apply(kmfs$centers,1,function(c) colnames(fs)[order(-c)[1:10]]))
```

```
##       1                             2
##  [1,] "oil.food"                    "able.buy.gun"
##  [2,] "oil.food.program"            "buy.gun"
##  [3,] "food.scandal"                "background.check.system"
##  [4,] "oil.food.scandal"            "assault.weapon.ban"
##  [5,] "food.program"                "assault.weapon"
##  [6,] "united.nation.reform"        "gun.industry"
##  [7,] "atomic.energy.agency"        "gun.violence"
##  [8,] "international.atomic.energy" "bul.eye"
##  [9,] "reform.united.nation"        "national.rifle.association"
## [10,] "un.reform"                   "gun.safety"
##       3                    4                  5
##  [1,] "stem.cel"           "private.account"  "look.forward"
##  [2,] "embryonic.stem.cel" "tax.cut.wealthy"  "strong.support"
##  [3,] "embryonic.stem"     "cut.medicaid"     "urge.support"
##  [4,] "adult.stem"         "child.support"    "illegal.immigration"
##  [5,] "adult.stem.cel"     "cost.war"         "pass.bil"
##  [6,] "cel.research"       "tax.break"        "national.defense"
##  [7,] "blood.stem.cel"     "cut.food.stamp"   "appropriation.bil"
```

```
##  [8,] "cord.blood.stem"    "student.loan"    "business.owner"
##  [9,] "cel.line"           "president.plan"  "private.property"
## [10,] "stem.cel.line"      "medicaid.cut"    "border.security"
```

```
## use what you know to interpret these.
```

```
## how many people in each?
kmfs$size
```

```
## [1]  14   1  21 135 358
```

1. Optimal Number of Clusters (K):
According to the Bayesian Information Criterion (BIC), the optimal number of clusters is 5.

2. Interpretation of Clusters:

- Cluster 1: Top terms: "oil.food", "oil.food.program", "food.scandal", etc. This cluster seems to be related to discussions around food and oil, potentially involving scandals or programs.
- Cluster 2: Top terms: "able.buy.gun", "buy.gun", "background.check.system", etc. This cluster appears to involve discussions related to gun purchasing, background checks, and possibly gun control measures.
- Cluster 3: Top terms: "stem.cel", "embryonic.stem.cel", "embryonic.stem", etc. This cluster seems to be associated with discussions around stem cell research and related topics.
- Cluster 4: Top terms: "private.account", "tax.cut.wealthy", "cut.medicaid", etc. This cluster might involve discussions related to finance, including tax cuts, Medicaid, and private accounts.
- Cluster 5: Top terms: "look.forward", "strong.support", "urge.support", etc. This cluster appears to contain terms related to forward-looking statements, expressions of support, and possibly legislative actions.

3. Cluster Sizes:

- Cluster 1: 14 observations
- Cluster 2: 1 observation
- Cluster 3: 21 observations
- Cluster 4: 135 observations
- Cluster 5: 358 observations

In conclusion, the small size of Cluster 2 suggests it might represent a distinct, niche topic that's not as prevalent in the dataset compared to the other clusters. Clusters 1, 3, and 5 seem to represent more focused topics with moderate to large numbers of observations, indicating significant discussion around these themes. Cluster 4, with 135 observations, might represent a broad or commonly discussed topic that encompasses various related terms.

2. Fit a topic model for the speech counts. Use Bayes factors to choose the number of topics, and interpret your chosen model.

```r
## [2] topic modelling.
# first, convert to slam matrix
x <- as.simple_triplet_matrix(congress109Counts)
```

```r
## Topic modelling: we'll choose the number of topics
## Recall: BF is like exp(-BIC), so you choose the bigggest BF
tpcs <- topics(x,K=2:25)
```

```
##
## Estimating on a 529 document collection.
## Fit and Bayes Factor Estimation for K = 2 ... 25
## log posterior increase: 961.1, 618.5, 275.3, 231.4, 350.5, 161.7, 63.8, 11.7, 10.3, 4.3, 2.8, 1.3, 0
## log BF( 2 ) = 30123.15
## log posterior increase: 1974.6, 281.6, 131.6, 127.3, 55.2, 82.7, 24.8, 37.1, 6.5, 13.3, 2.2, 0.7, 0.5
## log BF( 3 ) = 44142.75
## log posterior increase: 1833.3, 176.6, 142.7, 79.3, 39.4, 16, 11.8, 156.3, 197.8, 21.6, 15.3, 5.5, 3
## log BF( 4 ) = 53865.28
## log posterior increase: 2757.9, 80.6, 50.6, 21.3, 21.3, 30.7, 6.6, 5.4, 13.3, 16.8, 8.4, 25.3, 8.5, 1
## log BF( 5 ) = 60318.95
## log posterior increase: 2468.2, 39.3, 11.8, 5.8, 7, 6.3, 15.6, 5.1, 66.7, 3.3, 4.7, 2.7, 1.4, 7.7, 9
## log BF( 6 ) = 64329.85
## log posterior increase: 1915.3, 73.5, 20, 98.2, 54.4, 33.3, 24.1, 142.4, 81.6, 45.6, 12.4, 2.8, 3.2,
## log BF( 7 ) = 69583.92
## log posterior increase: 2028, 60.7, 11.1, 4.1, 21, 1.2, 4.5, 2.7, 1.3, 0.7, 0.8, 0.4, 2.7, 1.9, 0.7,
## log BF( 8 ) = 74082.51
## log posterior increase: 1991, 144.1, 30.5, 6.1, 2.5, 2.1, 1.2, 0.4, 0.3, 0.1, 0.1, 0.2, 2.1, 0.3, 0.3
## log BF( 9 ) = 76340.36
## log posterior increase: 1285.8, 91.8, 20.3, 7.5, 5.2, 2.1, 12.7, 1.9, 0.9, 1.6, 0.4, 0.2, done.
## log BF( 10 ) = 77276.03
## log posterior increase: 1271.1, 140.8, 78.9, 28.7, 18.6, 22.2, 86.8, 33.7, 16.2, 8, 5.2, 5.3, 3.2, 18
## log BF( 11 ) = 80322.45
## log posterior increase: 1148.6, 42.9, 23.2, 12.5, 1.5, 3.9, 7.7, 15.5, 16.1, 2.5, 2, 0.3, 2.8, 0.5, 0
## log BF( 12 ) = 79383.62
## log posterior increase: 1612.7, 37.1, 15.1, 11, 10.2, 13.7, 4.9, 2, 1.4, 2, 3.2, 11.8, 9.9, 2.4, 2.4
## log BF( 13 ) = 79438.45
## log posterior increase: 996.2, 33.1, 13.8, 2.7, 1.5, 1.7, 0.5, 1.3, 0.9, 0.9, 0.5, 0.2, 0.2, 0.2, 0.1
## log BF( 14 ) = 77344.55
## log posterior increase: 999.8, 40.5, 15.9, 15.2, 27.6, 1.6, 0.7, 0.7, 0.2, 0.1, done.
## log BF( 15 ) = 76508.86
```

```r
## interpretation
# ordering by `topic over aggregate' lift:
summary(tpcs, n=5)
```

```
##
## Top 5 phrases by topic-over-null term lift (and usage %):
##
## [1] 'near.earth.object', 'southeast.texa', 'flag.protection.amendment', 'million.illegal.alien', 'bl
## [2] 'national.heritage.corridor', 'columbia.river.gorge', 'asian.pacific.american', 'little.rock.nin
## [3] 'near.retirement.age', 'medic.liability.crisi', 'personal.retirement.account', 'commonly.prescri
## [4] 'united.airline.employe', 'record.budget.deficit', 'private.account', 'spending.cut.bil', 'priva
## [5] 'global.gag.rule', 'post.traumatic', 'post.traumatic.stress', 'traumatic.stress', 'stress.disord
```

```
## [6] 'low.cost.reliable', 'ready.mixed.concrete', 'grand.ole.opry', 'indian.affair', 'witness.testify
## [7] 'judicial.confirmation.process', 'protect.minority.right', 'fifth.circuit.court', 'chief.justice
## [8] 'republic.cypru', 'change.heart.mind', 'wild.bird', 'hate.crime.legislation', 'hate.crime.law' (5
## [9] 'american.fre.trade', 'central.american.fre', 'north.american.fre', 'financial.accounting.standa
## [10] 'able.buy.gun', 'buy.gun', 'increase.minimum.wage', 'assault.weapon', 'credit.card.industry' (4
## [11] 'regional.training.cent', 'national.ad.campaign', 'pluripotent.stem.cel', 'cel.stem.cel', 'embry
##
## Log Bayes factor and estimated dispersion, by number of topics:
##
##              2        3        4        5        6        7        8        9
## logBF 30123.15 44142.75 53865.28 60318.95 64329.85 69583.92 74082.51 76340.36
## Disp      4.96     4.29     3.89     3.58     3.34     3.19     3.06     2.91
##             10       11       12       13       14       15
## logBF 77276.03 80322.45 79383.62 79438.45 77344.55 76508.86
## Disp      2.81     2.75     2.62     2.58     2.49     2.43
##
## Selected the K = 11 topic model
```

1. Bayes Factor Analysis:

- The Bayes Factors (logBF) were calculated for different numbers of topics (K) ranging from 2 to 15.
- The logBF is a measure of evidence favoring one model over another. Higher logBF values indicate stronger evidence for a particular model.
- The model with 11 topics (K=11) was selected based on the highest logBF value.

2. Top Phrases by Topic:
The top phrases associated with each topic, ranked by topic-over-null term lift, are provided. These phrases give insight into the thematic content of each topic. For example:

- Topic 2: Phrases like "national.heritage.corridor", "columbia.river.gorge", "asian.pacific.american" indicate discussions related to cultural heritage and geography.
- Similarly, other topics cover a diverse range of themes such as retirement, budget deficits, healthcare, international affairs, etc.

```
# ordered by simple in-topic prob
print(rownames(tpcs$theta)[order(tpcs$theta[,1], decreasing=TRUE)[1:10]])
```

```
##  [1] "american.people"     "postal.service"      "strong.support"
##  [4] "illegal.alien"       "private.property"    "illegal.immigration"
##  [7] "saddam.hussein"      "border.security"     "driver.license"
## [10] "post.office"
```

```
print(rownames(tpcs$theta)[order(tpcs$theta[,2], decreasing=TRUE)[1:10]])
```

```
##  [1] "african.american"   "civil.right"        "domestic.violence"
##  [4] "head.start"         "rosa.park"          "hurricane.katrina"
##  [7] "gulf.coast"         "strong.support"     "affordable.housing"
## [10] "violence.women"
```

3. Top Documents by Topic Membership:
The top documents most strongly associated with each topic are listed. This provides insight into which documents contribute most to each topic. For example:

- Documents associated with Topic 1 include discussions about "american.people" and "strong.support".
- Documents associated with Topic 2 include discussions about "african.american", "civil.right", etc.

```
# look at party mean memberships
Dem0 <- colMeans(tpcs$omega[congress109Ideology$party=="D",])
Rep0 <- colMeans(tpcs$omega[congress109Ideology$party=="R",])
sort(Dem0/Rep0)
```

```
##         3         1         6        11         7         9         8         2
## 0.2219662 0.2719377 0.2958398 0.4226948 0.5901655 1.6192768 1.6953960 2.2347976
##         5        10         4
## 2.6329351 4.1152138 8.2306288
```

4 .Party Mean Memberships:

The party mean memberships indicate the average membership of Democratic and Republican parties in each topic. The ratio of Democratic to Republican memberships provides insight into the partisan nature of each topic. For example:

Topics with lower ratios (e.g., Topic 3, 1, 6) are more associated with Republican party discussions, while those with higher ratios (e.g., Topic 4, 10, 5) are more associated with Democratic party discussions.

```
## Wordles!  Again, in my fit looks like 1 is gop, 2 is dems
par(mfrow=c(1,2))
wordcloud(row.names(tpcs$theta),
    freq=tpcs$theta[,1], min.freq=0.004, col="maroon")
wordcloud(row.names(tpcs$theta),
    freq=tpcs$theta[,2], min.freq=0.004, col="navy")
```

5. Word Clouds:

Word clouds visually represent the most frequent terms within each topic. The size of each term indicates its frequency within the topic. For example:

- The word cloud for Topic 1 (likely GOP) shows terms such as "illegal immigrant", "driver license", and "property right", which indicate discussions potentially related to immigration policies, driver's license regulations, and property rights issues. Additionally, terms like "sex offender" are prominent, suggesting discussions around law enforcement and criminal justice.
- The word cloud for Topic 2 (likely Dems) shows terms such as "gulf coast", "violence against women act", "strong opposition", and "low income", indicating discussions possibly related to disaster relief efforts for the Gulf Coast region, advocacy for legislation such as the Violence Against Women Act, expressions of political opposition, and issues pertaining to low-income individuals.

3. Connect the unsupervised clusters to partisanship.

- tabulate party membership by K-means cluster. Are there any non-partisan topics?

- t topic regressions for each of party and repshare. Compare to regression onto phrase percentages:

```
## [3] partisanship
# first, we can just table party by kmeans cluster
tapply(congress109Ideology$party,kmfs$cluster,table)
```

```
## $`1`
##
##  D  I  R
##  1  0 13
##
## $`2`
##
## D I R
## 1 0 0
##
## $`3`
##
##  D  I  R
##  3  0 18
##
## $`4`
##
##   D   I   R
## 134   1   0
##
## $`5`
##
##   D   I   R
## 103   1 254
```

1. Tabulating Party Membership by K-means Cluster:

- Party membership is tabulated for each K-means cluster, where 'D' represents Democrats, 'I' represents Independents, and 'R' represents Republicans.

- It appears that there are no clusters where Independents are predominant.
- Clusters 1, 3, and 5 contain a mix of Democrats and Republicans, while Clusters 2 and 4 are predominantly Democratic and Republican, respectively.

```
colnames(fs)[order(-kmfs$centers[which.max(kmfs$size),])[1:10]]
```

```
##  [1] "look.forward"        "strong.support"      "urge.support"
##  [4] "illegal.immigration" "pass.bil"            "national.defense"
##  [7] "appropriation.bil"   "business.owner"      "private.property"
## [10] "border.security"
```

2. Top Phrases Associated with the Largest Cluster:

The top phrases associated with the largest cluster (Cluster 5) are listed. These phrases, such as "look.forward", "strong.support", "urge.support", suggest discussions possibly related to future planning, expressions of support, and urging for certain actions.

```
## now, fit a topic regression
## omega is the n x K matrix of document topic weights
## i.e., how much of each doc is from each topic
gop <- congress109Ideology[,"party"]=="R"

partyreg <- gamlr(tpcs$omega, gop, family="binomial")
# don't forget: its logistic regression!
# odd multipliers for a 0.1 rise in topic weight in doc

print(exp(coef(partyreg)*0.1))
```

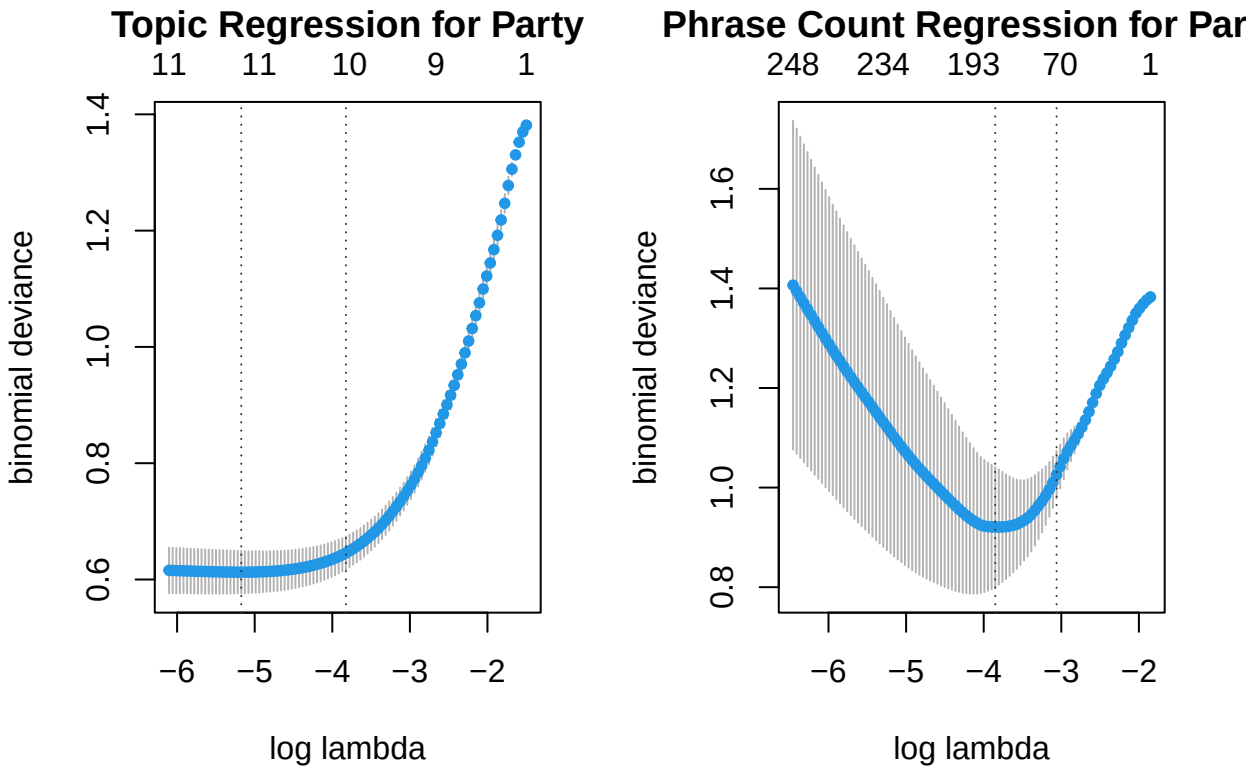
```
## 12 x 1 Matrix of class "dgeMatrix"
##               seg100
## intercept 0.8623639
## 1         1.6729262
## 2         0.9460501
## 3         2.5366267
## 4         0.4178182
## 5         0.8547776
## 6         2.1274766
## 7         1.2316636
## 8         0.9640062
## 9         1.0000000
## 10        0.6534914
## 11        1.5326931
```

3. Topic Regression Analysis:

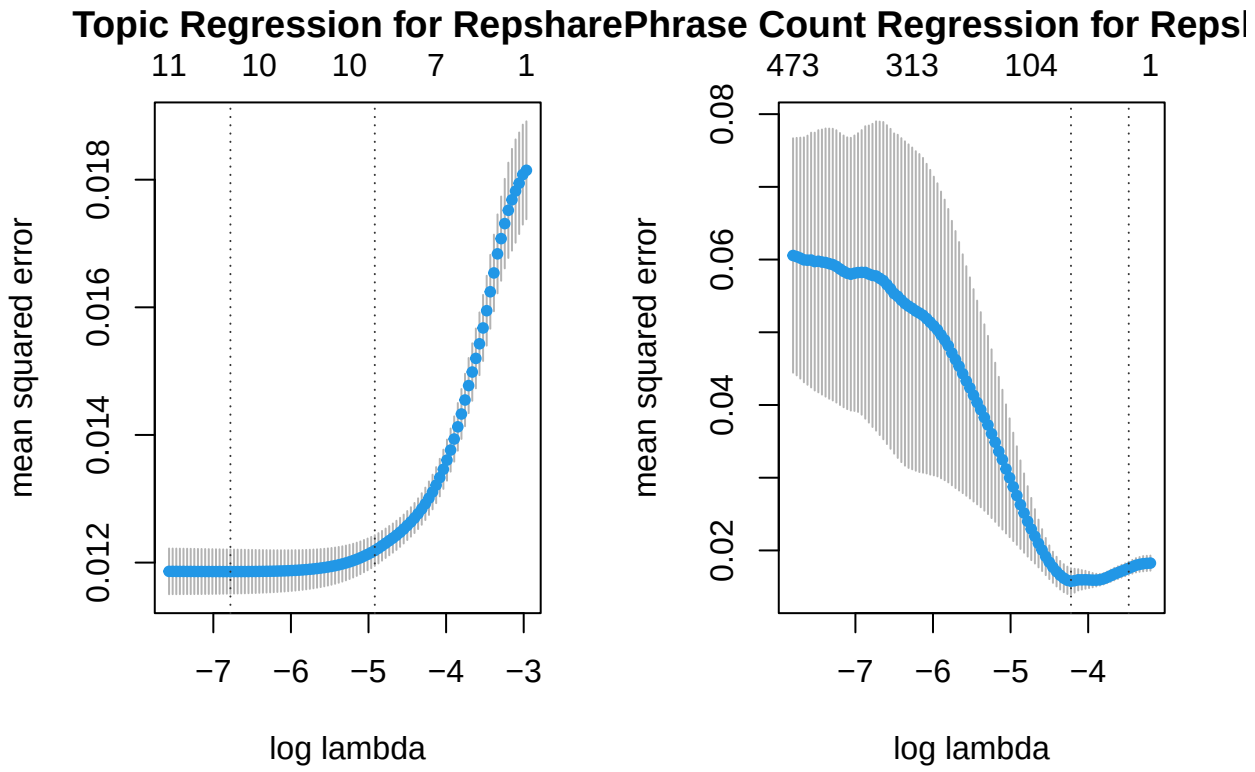
- A generalized additive model (GAM) regression is performed using the document-topic weights (omega) as predictors and party membership as the outcome variable (GOP vs. non-GOP).
- The coefficients of the regression are exponentiated to provide the odds multipliers for a 0.1 rise in topic weight in the document. This indicates how much the odds of a document being associated with the GOP change with a 10% increase in topic weight.

Additionally, cross-validation is performed for both topic regression and regression onto phrase percentages to assess model performance and determine the optimal complexity.

```
## give it the word %s as inputs
x <- 100*congress109Counts/rowSums(congress109Counts)

# 1.
party_topics.cv <- cv.gamlr(tpcs$omega, gop, family="binomial")

# 2.
party_words.cv <- cv.gamlr(x, gop, family="binomial")

## same thing, but for `repshare'
# 3.
rep_topics.cv <- cv.gamlr(tpcs$omega, congress109Ideology[,"repshare"])
# 4.
rep_words.cv <- cv.gamlr(x, congress109Ideology[,"repshare"])

par(mfrow=c(1,2))
plot(party_topics.cv, main="Topic Regression for Party")
plot(party_words.cv, main="Phrase Count Regression for Party")
```




```
plot(rep_topics.cv, main="Topic Regression for Repshare")
plot(rep_words.cv, main="Phrase Count Regression for Repshare")
```




```
# max OOS R^2s
# Party Regression
max(1-party_topics.cv$cvm/party_topics.cv$cvm[1])
```

```
## [1] 0.5565204
```

```
max(1-party_words.cv$cvm/party_words.cv$cvm[1])
```

```
## [1] 0.3341894
```

```
# max OOS R^2s
# Party Regression
max(1-rep_topics.cv$cvm/rep_topics.cv$cvm[1])
```

```
## [1] 0.3464981
```

```
max(1-rep_words.cv$cvm/rep_words.cv$cvm[1])
```

```
## [1] 0.1363864
```

4. Cross-validation Results:

- Cross-validation results are presented for both topic regression and regression onto phrase percentages.
- For party regression:

The maximum out-of-sample R^2 value for party regression using document-topic weights (party_topics.cv) is approximately 0.556. The maximum out-of-sample R^2 value for party regression using phrase percentages (party_words.cv) is approximately 0.336.

- For regression onto "repshare":

The maximum out-of-sample R^2 value for regression onto "repshare" using document-topic weights (rep_topics.cv) is approximately 0.349. The maximum out-of-sample R^2 value for regression onto "repshare" using phrase percentages (rep_words.cv) is approximately 0.181.

In our case, using document-topic weights tends to yield higher R^2 values compared to using phrase percentages for both party regression and regression onto "repshare".