12 May 2015

A Logistic Regression model of the changing English preterit

Research results - Methodology & Statistics for Linguistc Research

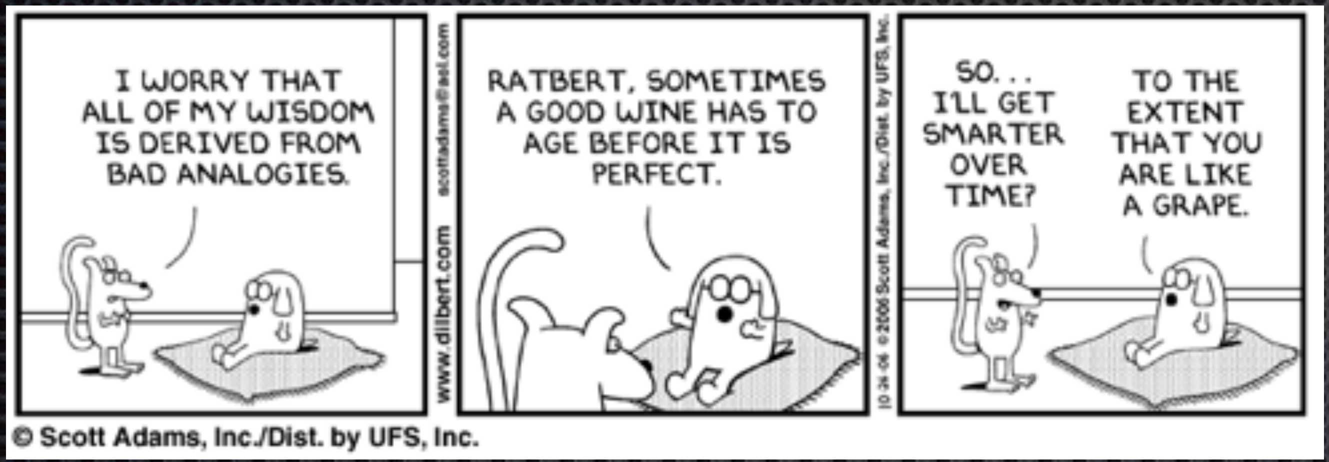
Esther van den Berg

- Introduction
- Background
  - Analogical Modeling
  - Regularisation
- Method
  - Collecting Data
  - First impression of Data
  - Logistic Regression
  - Diagnostics
- Results
- Discussion
- Conclusion

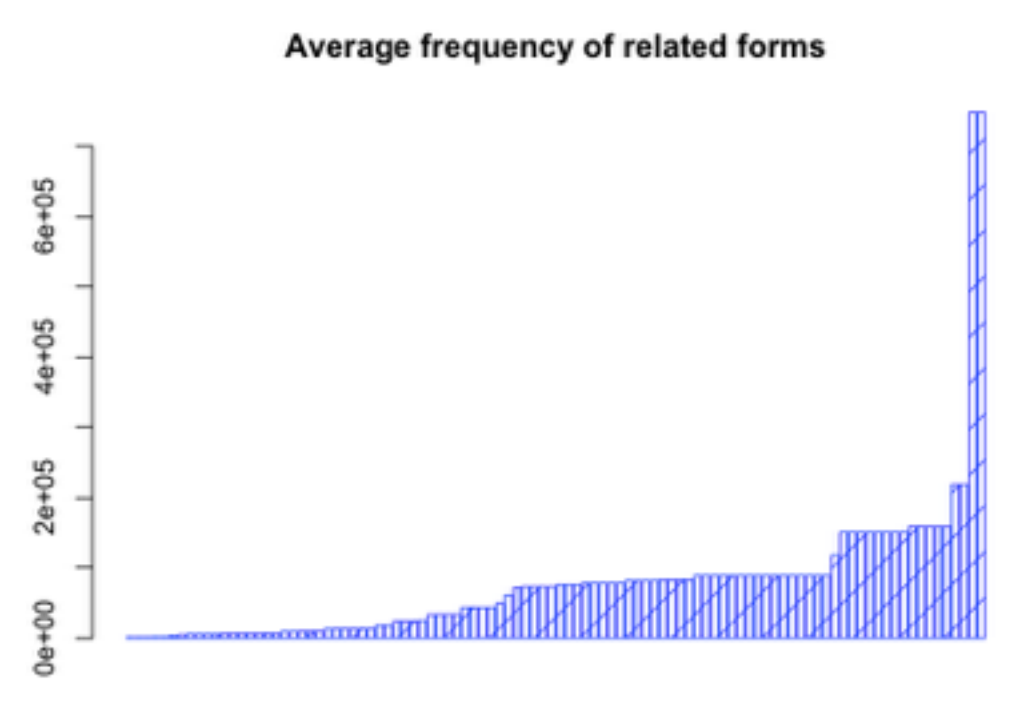
Introduction

- Two notions to keep in mind

1) Analogy as a model for language



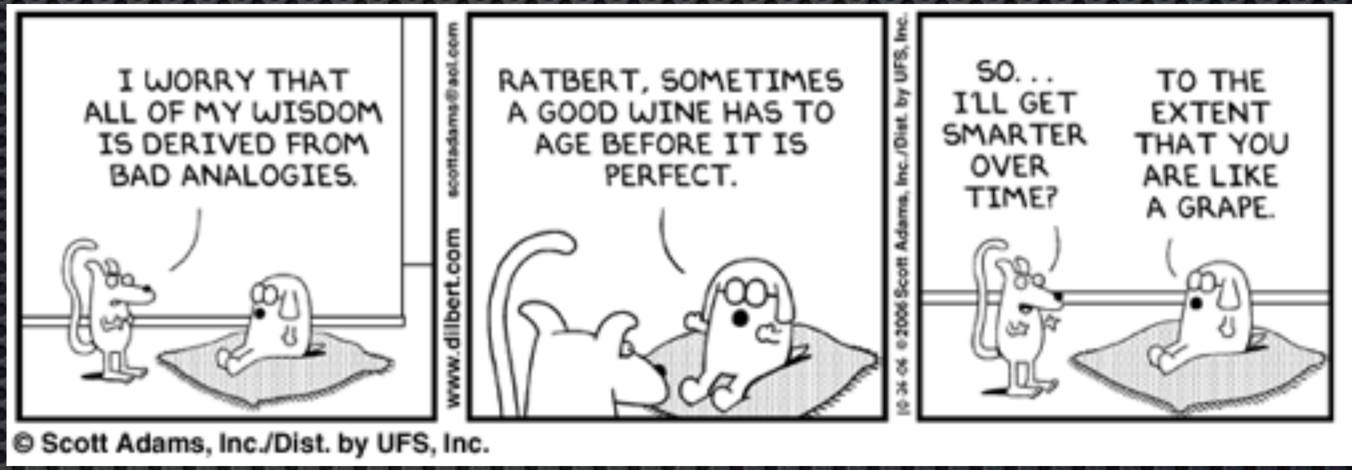
2) Dealing with frequency data



Introduction

"A comparison between one thing and another, typically for the purpose of explanation or clarification"

- Oxford Dictionary



Introduction

- analogical processes
  - forms may change class because they resemble other forms

present	praeterit
grow	grew
claw	clew
saw	sawed

→

present	praeterit
grow	grew
claw	clawed
saw	sawed

Introduction

- analogical processes
  - forms may change class because they resemble other forms
- frequency effects
  - “irregular” form may persist because of its frequency

present	**praeterit**
grow	grew
claw	clew
saw	sawed

present	**praeterit**
grow	**grew**
claw	**clawed**
saw	sawed

Introduction

- analogical processes
    - forms may change class because they resemble other forms
- frequency effects
    - “irregular” form may persist because of its frequency

present	praeterit
grow	grew
claw	clew
saw	sawed

present	praeterit
grow	**grew**
claw	**clawed**
saw	sawed

Stable vs Changeable Items

Introduction

No research has been done to determine whether frequent and infrequent forms are equally likely to be used as a basis for analogy

- a form’s stability could depend on the presence of a group of frequent, analogous words
- or it could depend on the presence of any single frequent analogous form

Introduction

No research has been done to determine whether frequent and infrequent forms are equally likely to be used as a basis for analogy

- a form's stability could depend on the presence of a group of frequent, analogous words
- or it could depend on the presence of any single frequent analogous form

1. Is the stability of English strong verbs influenced by the average frequency of its analogically related forms?

2. Is the stability of English strong verbs influenced by the maximally frequent form of its analogically related forms?

Analogical Modeling

- Simulating linguistic behavior by assuming the presence of analogy in linguistic representations and treating linguistic structures as (potential) analogical concepts
- A structure can function analogically if inserting items into that structure guarantees similarity of meaning

Analogical Modeling

- Simulating linguistic behavior by assuming the presence of analogy in linguistic representations and treating linguistic structures as (potential) analogical concepts
- A structure can function analogically if inserting items into that structure guarantees similarity of meaning
- Often used to provide an explanation for morphological developments

Analogical Modeling

- Simulating linguistic behavior by assuming the presence of analogy in linguistic representations and treating linguistic structures as (potential) analogical concepts
- A structure can function analogically if inserting items into that structure guarantees similarity of meaning
- Often used to provide an explanation for morphological developments

present	past tense
drive	drove
ride	rode
strive	strove

Analogical Modeling

- Simulating linguistic behavior by assuming the presence of analogy in linguistic representations and treating linguistic structures as (potential) analogical concepts
- A structure can function analogically if inserting items into that structure guarantees similarity of meaning
- Often used to provide an explanation for morphological developments

present	**past tense**
drive	drove
ride	rode
strive	strove
dive	dove

Analogical Modeling

present	praeterit
grow	grew
claw	clew
saw	sawed

present	praeterit
grow	**grew**
claw	**clawed**
saw	sawed

Stable vs Changeable Items

More commonly, strong verbs become weak --> ***regularisation***

Analogical Modeling

- Albright & Hayes, 2002
  - development of **Minimal Generalisation learner** as an automated analogous predictor
  - generalizes from word-specific rules to derive analogous patterns
- Krygier 1994
  - Overview of English strong verb system and the various factors which played a role in the disappearance of many strong forms

Method

- Collecting Data
- First impression of Data
- Logistic Regression
- Diagnostics

Method

- Collecting Data
  - 100 verbs and their preterit form in Middle English (ME) and Modern English (ModE) from Krygier 1994
  - Note for each their status as either **stable** or **changed**
  - Fed to Albright & Hayes' Minimal Generalization Learner to obtain analogical forms

Method

- Collecting Data
  - 100 verbs and their preterit form in Middle English (ME) and Modern English (ModE) from Krygier 1994
  - Note for each their status as either **stable** or **changed**
  - Fed to Albright & Hayes' Minimal Generalization Learner to obtain analogical forms
- From the output —> average and maximum frequency of related forms
  - Dependent variable: categorical
  - Independent variable: continous

Method

- Collecting Data
  - 100 verbs and their preterit form in Middle English (ME) and Modern English (ModE) from Krygier 1994
  - Note for each their status as either **stable** or **changed**
  - Fed to Albright & Hayes' Minimal Generalization Learner to obtain analogical forms
- From the output —> average and maximum frequency of related forms
  - Dependent variable: categorical
  - Independent variable: continous

-> Logistic regression

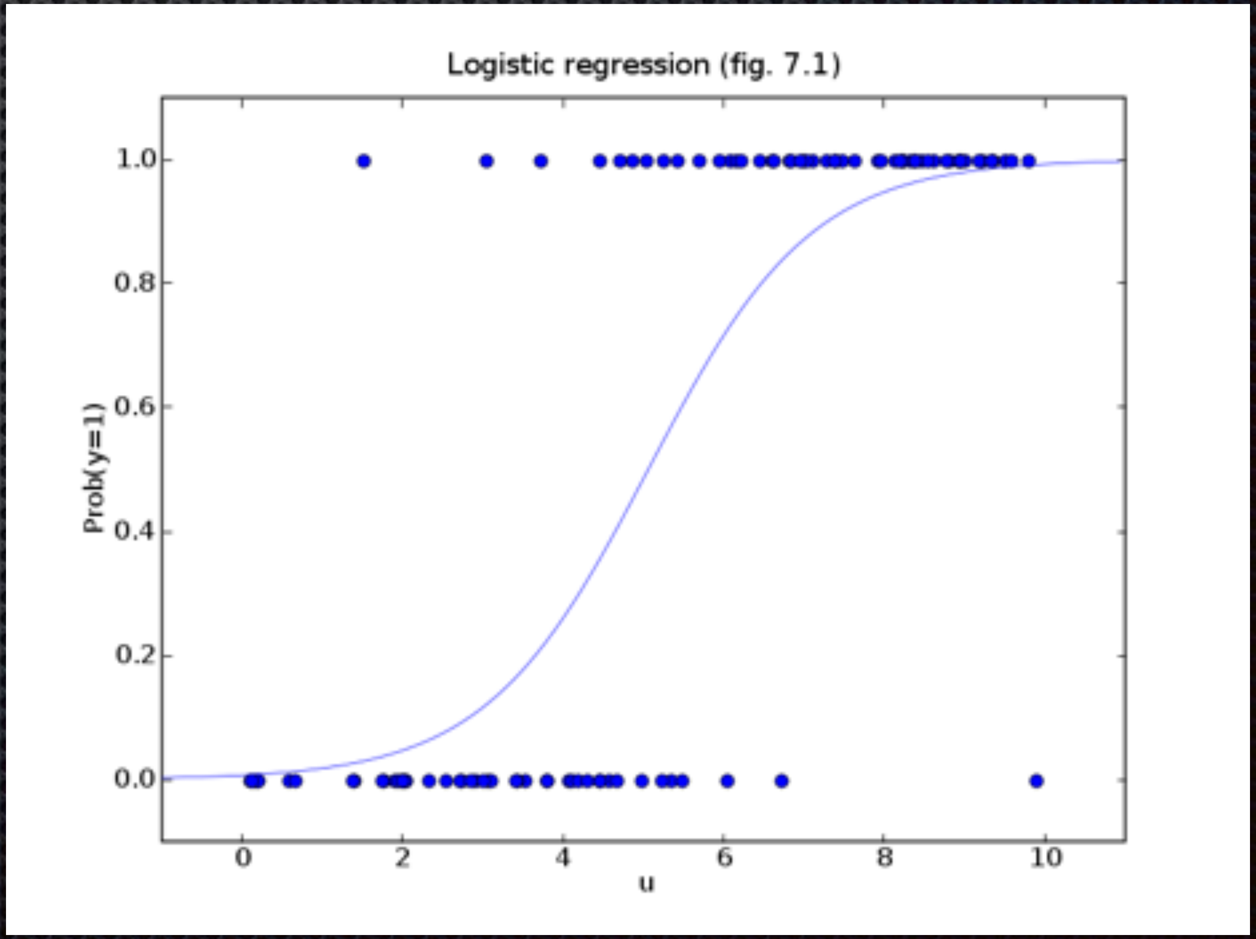
LR

- Logistic regression
  - Maximum Likelihood Estimation: making the model's prediction most similar to the observed data
  - LINK function to express binary variable as probabilities
  - Log odds ratio $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$

LR

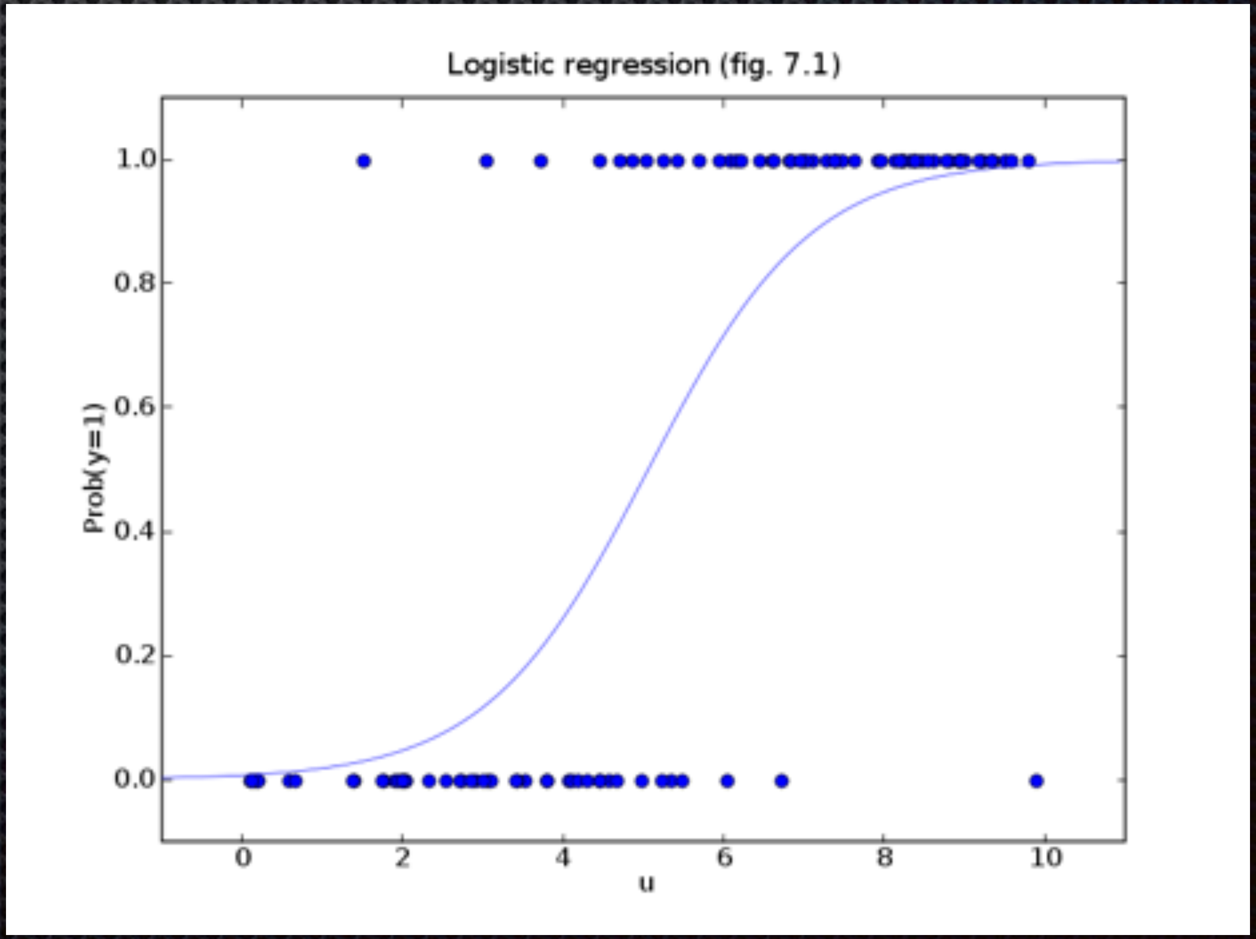
- Logistic regression
  - Maximum Likelihood Estimation: making the model's prediction most similar to the observed data
  - LINK function to express binary variable as probabilities
  - Log odds ratio

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$$



LR

- Logistic regression
  - Maximum Likelihood Estimation: making the model's prediction most similar to the observed data
  - LINK function to express binary variable as probabilities
  - Log odds ratio $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$
- In R:
  - Specify a model to be fit to the data by means of a formula



LR

- Deviance residuals
  - similar to difference between observed and expected values
- Coefficients
  - Negative coefficients indicate that the chance of a correct response goes down
- Residual deviance to check for overdispersion

Assumptions

- No overfitting or underfitting: include only and all meaningful variables
- Independent variables and log odds should be linearly related
- Large sample sizes

Variables

- token frequency - www.wordandphrase.info/
- average frequency
- maximum frequency
- type frequency

Variables

- token frequency - www.wordandphrase.info/
- average frequency
- maximum frequency
- type frequency

predicting

- status (**stable** or **changed**)

Variables

- token frequency - www.wordandphrase.info/
- average frequency
- maximum frequency
- type frequency

predicting

- status (**stable** or **changed**)

form	correct	ranking	missing	confidence	token freq	average freq	max freq	type freq
1~100	TRUE or FALSE	1~3	TRUE or FALSE	0~1	corpus counts	~80 000	~200000	1~12

Variables

- token frequency - www.wordandphrase.info/
- average frequency
- maximum frequency
- type frequency

predicting

- status (**stable** or **changed**)

form	correct	~~ranking~~	~~missing~~	confidence	token freq	average freq	max freq	type freq
1~100	TRUE or FALSE	~~1~3~~	~~TRUE or FALSE~~	0~1	corpus counts	~80 000	~200000	1~12

Variables

- token frequency - www.wordandphrase.info/
- average frequency
- maximum frequency
- type frequency

predicting

- status (**stable** or **changed**)

form	correct	ranking	missing	confidence	token freq	average freq	max freq	type freq
1~100	TRUE or FALSE	1~3	TRUE or FALSE	0~1	corpus counts	~80 000	~200000	1~12

Method

- token frequency - www.wordandphrase.info/
- average frequency
- maximum frequency
- type frequency

form	pattern	form1		form2		A		B	Change		Pres	P	scope	hits	reliability	confidence	related forms	exceptions
14	14	b1nd	->	b2nd	by	1	->	2	1/2/2	/	X		4	3	0.75	0.525862123591678	b1nd, f1nd, w1nd	bl1nd
14	4	b1nd	-	b1ndld	by	[]	-	ld	/ld/0	/	X	d	45	21	0.46666666666	0.41681499949	ad, av3d, bl1nd, bOrd	b1nd, bEnd, bld, bld, blld, brid, dlv1d

Method

- token frequency - www.wordandphrase.info/
- average frequency
- maximum frequency
- type frequency

form	pattern	form1		form2		A		B	Change		Pres	P	scope	hits	reliability	confidence	related forms	exceptions
14	14	b1nd	->	b2nd	by	1	->	2	1/2/2	/	X		4	3	0.75	0.525862123591678	b1nd, f1nd, w1nd	bl1nd
14	4	b1nd	-	b1ndld	by	[]	-	ld	/ld/0	/	X	d	45	21	0.46666666666	0.41681499949	ad, av3d, bl1nd, bOrd,	b1nd, bEnd, bld, bld, blld, brid, dlv1d

Method

- token frequency - www.wordandphrase.info/
- average frequency
- maximum frequency
- type frequency

token frequency

form	pattern	form1		form2		A		B	Change		Pres	P	scope	hits	reliability	confidence	related forms	exceptions
14	14	b1nd	->	b2nd	by	1	->	2	1/2/2	/	X		4	3	0.75	0.525862123591678	b1nd, f1nd, w1nd	bl1nd
14	4	b1nd	-	b1ndld	by	[]	-	ld	/ld/0	/	X	d	45	21	0.46666666666	0.41681499949	ad, av3d, bl1nd, bOrd,	b1nd, bEnd, bld, bld, blld, brid, dlv1d,

Method

- token frequency - www.wordandphrase.info/
- average frequency
- maximum frequency
- type frequency

- sum of class members = type frequency

- sum of token frequencies / type frequency = average frequency

- max of frequencies

token frequency

for m	patter n	form1		form2		A		B	Chang e		Pre s	P	scop e	hit s	reliability	confidence	related forms	exceptions
14	14	b1nd	->	b2nd	by	1	->	2	1/2/2	/	X		4	3	0.75	0.525862123591678	b1nd, f1nd, w1nd	bl1nd
14	4	b1nd	-	b1ndId	by	[]	-	Id	/Id/0	/	X	d	45	21	0.46666666666	0.41681499949	ad, av3d, bl1nd, bOrd	b1nd, bEnd, bld, bld, blld, brid, dlv1d

Method

- token frequency - www.wordandphrase.info/
- average frequency
- maximum frequency
- type frequency

- sum of class members = type frequency

- sum of token frequencies / type frequency = average frequency

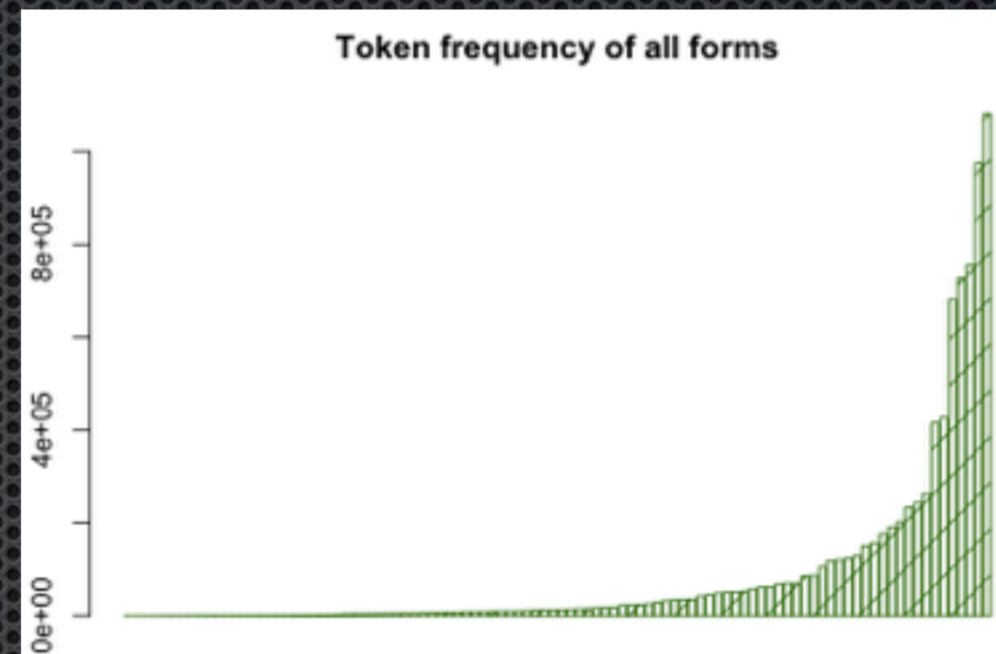
- max of frequencies

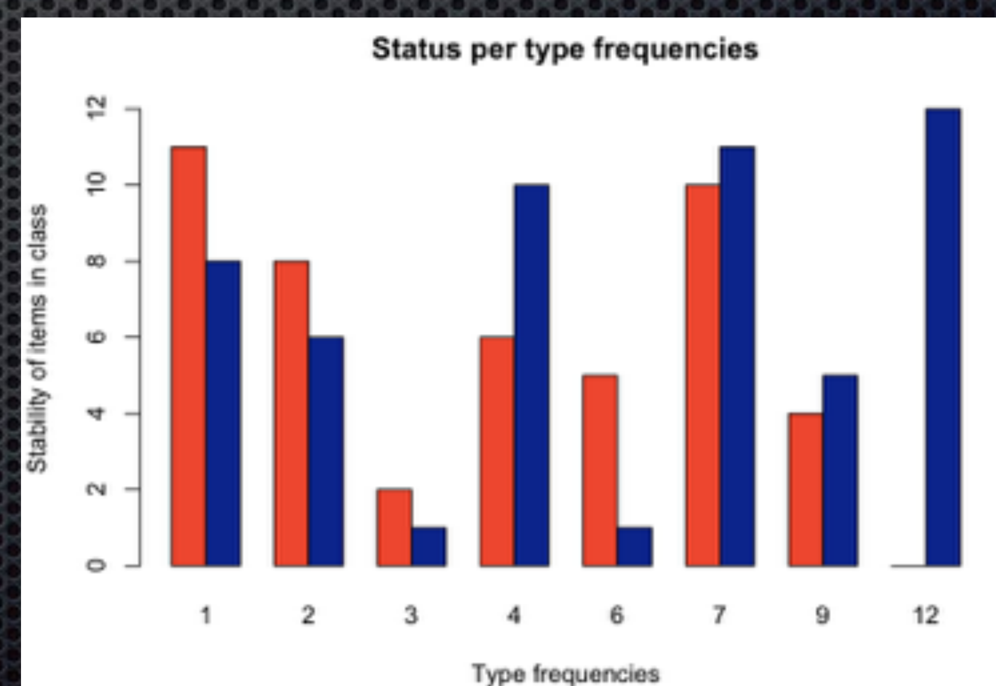
token frequency

form	pattern	form1		form2		A		B	Change		Pres	P	scope	hits	reliability	confidence	related forms	exceptions
14	14	b1nd	->	b2nd	by	1	->	2	1/2/2	/	X		4	3	0.75	0.525862123591678	b1nd, f1nd, w1nd	bl1nd
14	4	b1nd	-	b1ndld	by	[]	-	ld	/ld/0	/	X	d	45	21	0.46666666666	0.41681499949	ad, av3d, bl1nd, bOrd,	b1nd, bEnd, bld, bld, blld, brid, dly1d

First impression of Data

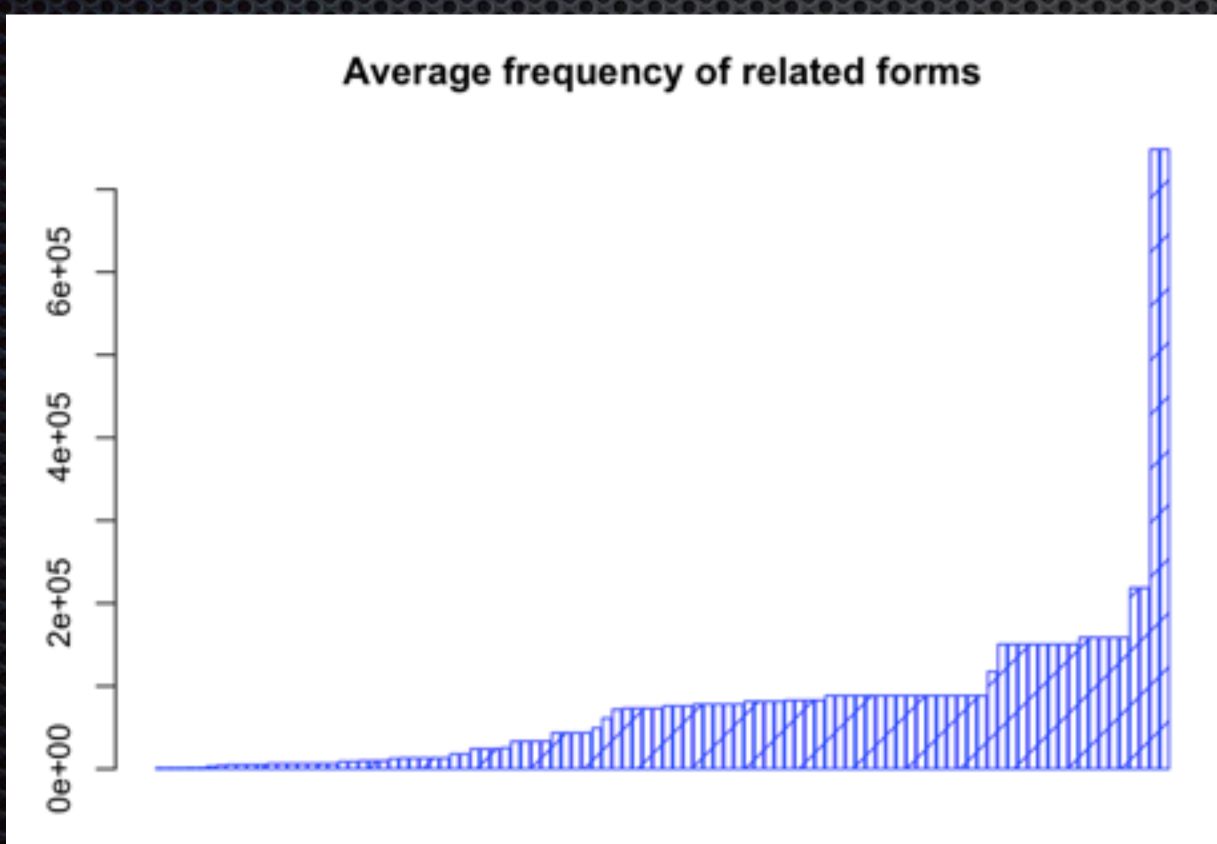
```
Phonological Learner File
Adam Albright/Bruce Hayes
English
Pretend Languages
Morphological categories:
Present Past
Input forms:
1s      1st
ad      adId
adjust  adjustId
admIt   admItId
adrEs   adrEst
aksEpt  aksEptId
aksEs   aksEst
akt     aktId
aNgyr   aNgyrd
ansyr   ansyrd
asUm    asUmd
av3d    av3dId
b1      b0t
b1nd    b2nd
b1t     bIt
bat     batId
batyl   batyld
bek     bekt
bEnd    bEnt
bEr     bor
bes     best
```

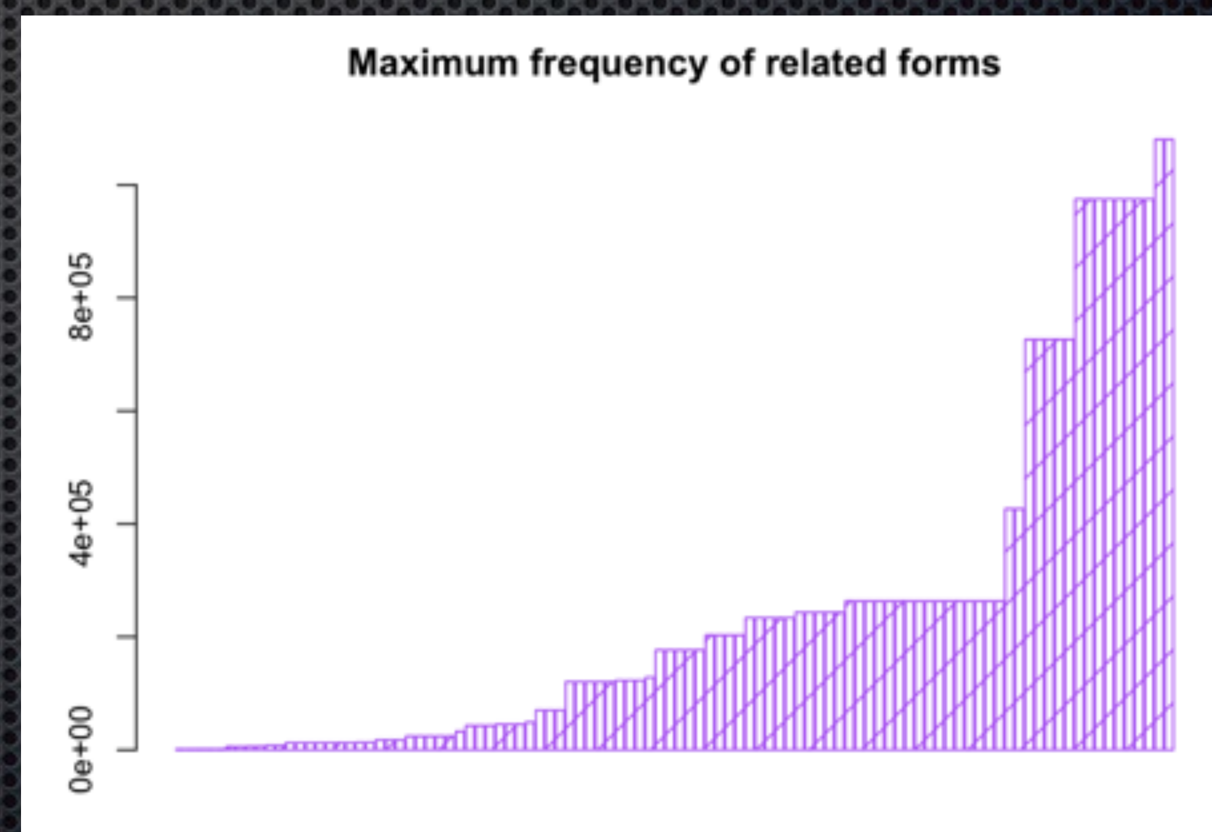




First impression of Data

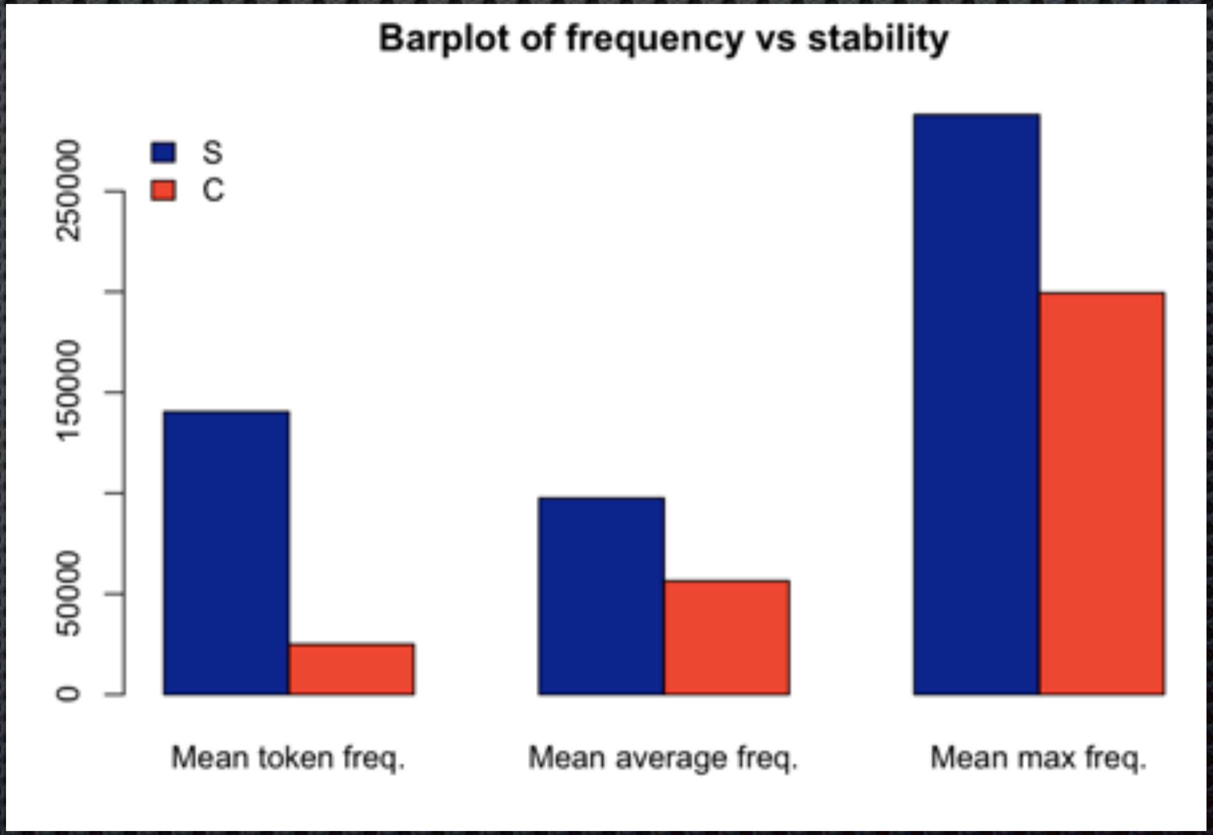
form	pattern	form1		form2		A		B	Change		Pres	P	scope	hits	reliability	confidence	related forms	exceptions
14	14	b1nd	->	b2nd	by	1	->	2	1/2/2	/	X		4	3	0.75	0.525862123591678	b1nd, f1nd, w1nd	bl1nd
14	4	b1nd	-	b1ndld	by	[]	-	ld	/ld/0	/	X	d	45	21	0.46666666666	0.41681499949	ad, av3d, bl1nd, bOrd	b1nd, bEnd, bld, bld, blld, brid, dlv1d



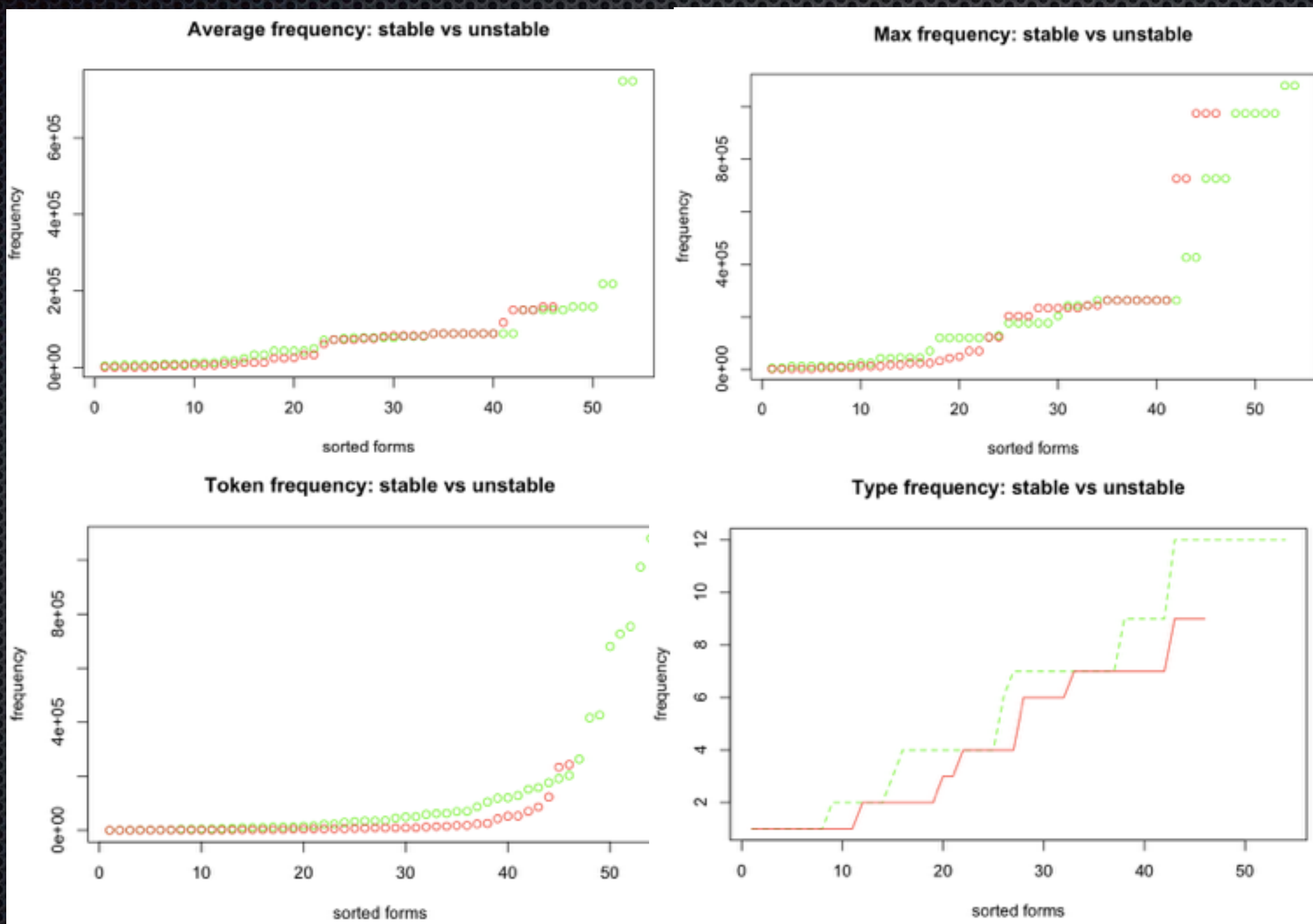


First impression of Data

for m	patter n	form1		form2		A		B	Chang e		Pre s	P	scop e	hit s	reliability	confidence	related forms	exceptions
14	14	b1nd	->	b2nd	by	1	->	2	1/2/2	/	X		4	3	0.75	0.525862123591678	b1nd, f1nd, w1nd	bl1nd
14	4	b1nd	->	b1ndld	by	[]	->	ld	/ld/0	/	X	d	45	21	0.4666666666666667	0.4168149994917332	ad, av3d, bl1nd, bOrd, dimand, dipEnd, End, fOld, g1d, hEd, lnklUd, kArd, nid, p2nd, pritEnd, s2nd, suksid, syr2nd, tred	b1nd, bEnd, bld, bld, blld, brid, dlv1d, f1nd, fid, h1d, hOld, lEnd, lid, r1d, rid, rld, sEnd, sl1d, spEnd



First impression of Data



Logistic regression

```
Call:
glm(formula = status ~ tokfreq + maxfreq + avfreq + typfreq,
    family = binomial, data = OEV)
```

Logistic regression

```
Call:
glm(formula = status ~ tokfreq + maxfreq + avfreq + typfreq,
    family = binomial, data = OEV)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.38296  -0.85382   0.03042   0.89985   1.74047

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.004e+00  6.282e-01  -3.190 0.001423 **
tokfreq      1.232e-05  4.782e-06   2.577 0.009971 **
maxfreq     -1.821e-06  2.011e-06  -0.905 0.365309
avfreq       8.086e-06  1.124e-05   0.719 0.472000
typfreq      2.763e-01  8.030e-02   3.440 0.000581 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137.99  on 99  degrees of freedom
Residual deviance: 108.16  on 95  degrees of freedom
AIC: 118.16

Number of Fisher Scoring iterations: 6
```

Logistic regression

```
Call:
glm(formula = status ~ tokfreq + maxfreq + avfreq + typfreq,
    family = binomial, data = OEV)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.38296  -0.85382   0.03042   0.89985   1.74047

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.004e+00  6.282e-01  -3.190 0.001423 **
tokfreq      1.232e-05  4.782e-06   2.577 0.009971 **
maxfreq     -1.821e-06  2.011e-06  -0.905 0.365309
avfreq       8.086e-06  1.124e-05   0.719 0.472000
typfreq      2.763e-01  8.030e-02   3.440 0.000581 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137.99  on 99  degrees of freedom
Residual deviance: 108.16  on 95  degrees of freedom
AIC: 118.16

Number of Fisher Scoring iterations: 6
```

Logistic regression

```
Call:
glm(formula = status ~ tokfreq + maxfreq + avfreq + typfreq,
    family = binomial, data = OEV)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.38296  -0.85382   0.03042   0.89985   1.74047

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.004e+00  6.282e-01  -3.190 0.001423 **
tokfreq      1.232e-05  4.782e-06   2.577 0.009971 **
maxfreq     -1.821e-06  2.011e-06  -0.905 0.365309
avfreq       8.086e-06  1.124e-05   0.719 0.472000
typfreq      2.763e-01  8.030e-02   3.440 0.000581 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137.99  on 99  degrees of freedom
Residual deviance: 108.16  on 95  degrees of freedom
AIC: 118.16

Number of Fisher Scoring iterations: 6
```

	2.5 %	97.5 %
token freq	1.00000423	1.0000229
max freq	0.99999404	1.0000017
aver. freq	0.99999133	1.0000017
type freq.	1.13813048	1.5642889

Exponentiated coefficients:

(Intercept)	tokfreq	maxfreq	avfreq	typfreq
0.1348249	1.0000123	0.9999982	1.0000081	1.3182270

Logistic regression

```
Call:
glm(formula = status ~ tokfreq + maxfreq + avfreq + typfreq,
    family = binomial, data = OEV)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.38296  -0.85382   0.03042   0.89985   1.74047

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.004e+00  6.282e-01  -3.190 0.001423 **
tokfreq      1.232e-05  4.782e-06   2.577 0.009971 **
maxfreq     -1.821e-06  2.011e-06  -0.905 0.365309
avfreq       8.086e-06  1.124e-05   0.719 0.472000
typfreq      2.763e-01  8.030e-02   3.440 0.000581 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137.99  on 99  degrees of freedom
Residual deviance: 108.16  on 95  degrees of freedom
AIC: 118.16

Number of Fisher Scoring iterations: 6
```

	2.5 %	97.5 %
token freq	*1.00000423*	*1.0000229*
max freq	~~0.99999404~~	~~1.0000017~~
aver. freq	~~0.99999133~~	~~1.0000017~~
type freq.	1.13813048	1.5642889

Exponentiated coefficients:

(Intercept)	tokfreq	maxfreq	avfreq	typfreq
0.1348249	1.0000123	0.9999982	1.0000081	1.3182270

Logistic regression

```
Call:
glm(formula = status ~ tokfreq + maxfreq + avfreq + typfreq,
    family = binomial, data = OEV)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.38296  -0.85382   0.03042   0.89985   1.74047

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.004e+00  6.282e-01  -3.190 0.001423 **
tokfreq      1.232e-05  4.782e-06   2.577 0.009971 **
maxfreq     -1.821e-06  2.011e-06  -0.905 0.365309
avfreq       8.086e-06  1.124e-05   0.719 0.472000
typfreq      2.763e-01  8.030e-02   3.440 0.000581 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137.99  on 99  degrees of freedom
Residual deviance: 108.16  on 95  degrees of freedom
AIC: 118.16

Number of Fisher Scoring iterations: 6
```

Exponentiated coefficients:

(Intercept)	tokfreq	maxfreq	avfreq	typfreq
0.1348249	1.0000123	0.9999982	1.0000081	1.3182270

- Multiple logistic regression shows that the model makes better predictions
- But only the effect of "token frequency" and "type frequency" was significant (β = 1.23, $p < .005$ and β = 2.76, $p < .001$)
- We cannot reject the null-hypothesis that the frequency of unrelated forms do not contribute to a stable outcome

Logistic regression

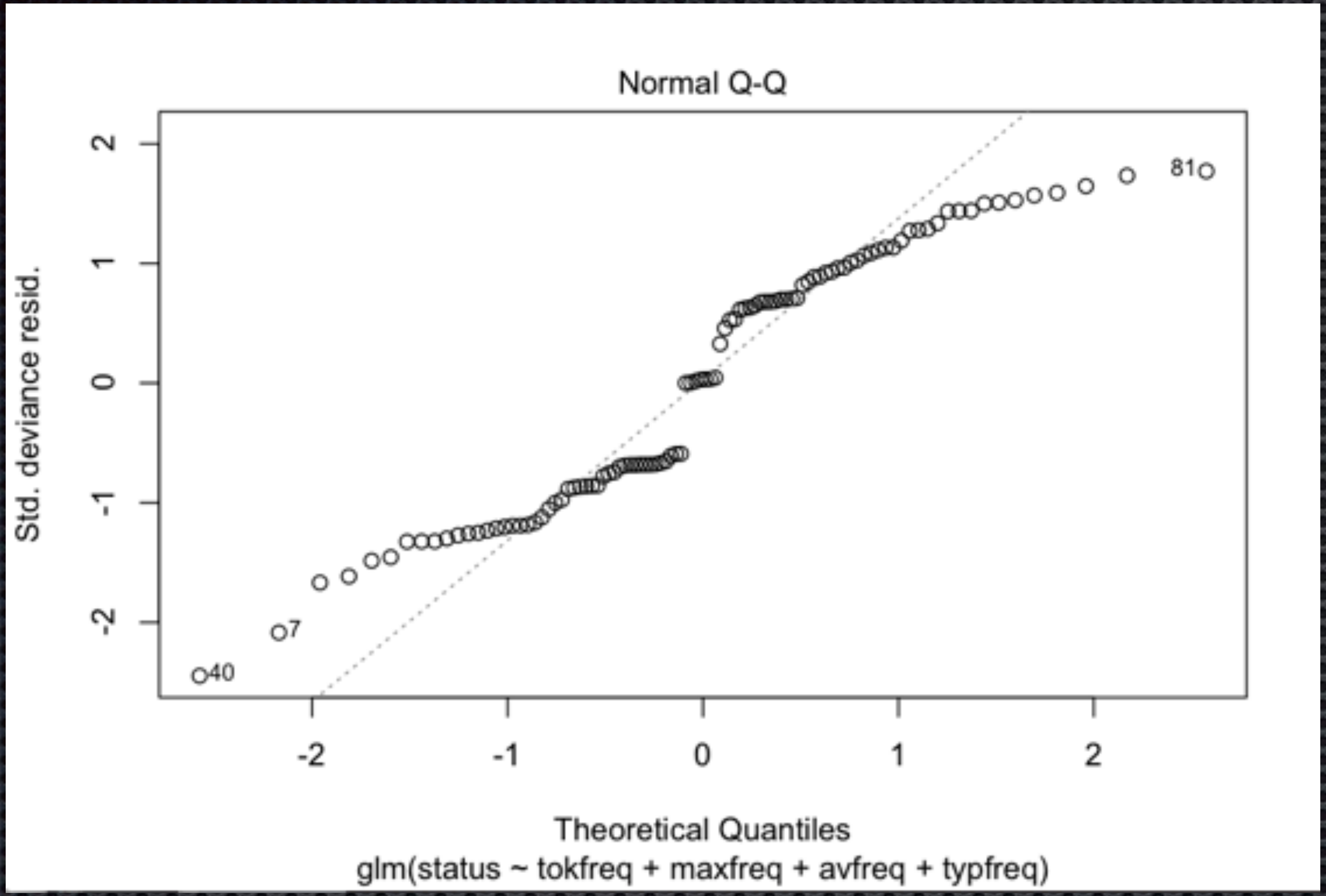
```
Call:
glm(formula = status ~ tokfreq + maxfreq + avfreq + typfreq,
    family = binomial, data = OEV)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.38296  -0.85382   0.03042   0.89985   1.74047

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.004e+00  6.282e-01  -3.190 0.001423 **
tokfreq      1.232e-05  4.782e-06   2.577 0.009971 **
maxfreq     -1.821e-06  2.011e-06  -0.905 0.365309
avfreq       8.086e-06  1.124e-05   0.719 0.472000
typfreq      2.763e-01  8.030e-02   3.440 0.000581 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137.99  on 99  degrees of freedom
Residual deviance: 108.16  on 95  degrees of freedom
AIC: 118.16

Number of Fisher Scoring iterations: 6
```

Exponentiated coefficients:

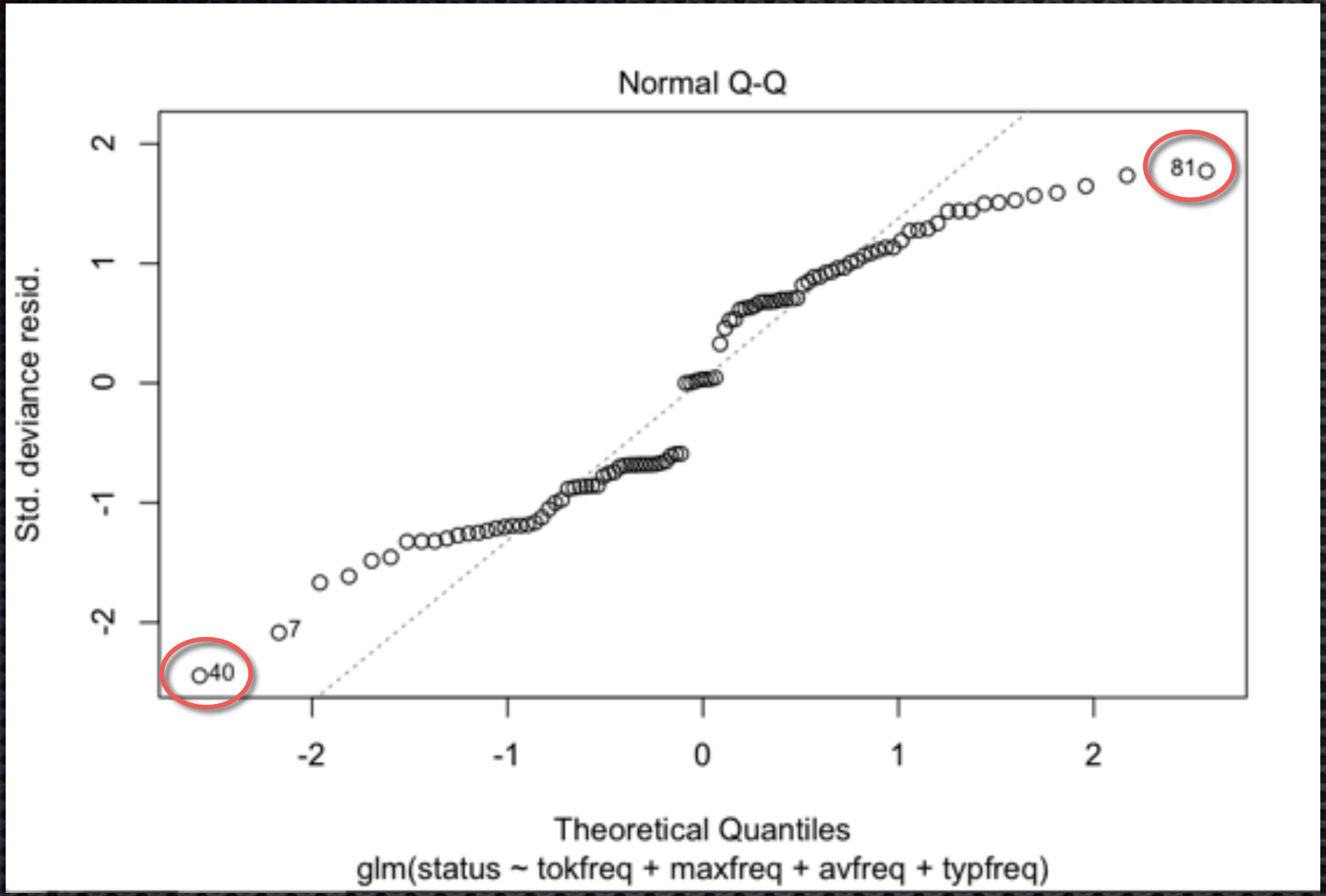
(Intercept)	tokfreq	maxfreq	avfreq	typfreq
0.1348249	1.0000123	0.9999982	1.0000081	1.3182270

- Multiple logistic regressions shows that the model makes better predictions
- But only the effect of "token frequency" and "type frequency" was significant ($\beta = 1.23$, $p < .005$ and $\beta = 2.76$, $p < .001$)
- We cannot reject the null-hypothesis that the frequency of related forms do not contribute to a stable outcome

Logistic regression



Logistic regression



- Originally:

Null deviance: 137.99
Residual deviance: 108.16
AIC: 118

- Without these outliers:

Null deviance: 135.203
Residual deviance: 98.137
AIC: 100.46

Logistic regression

- goodness of fit:
- “The question of how much better the model predicts the outcome variable can be assessed using the model chi-square statistic, which measures the difference between the model as it currently stands and the model when only the constant was included. ” (Field)
- 1 - pchisq(difference_in_deviance, difference_in_df) —> **0.0000052948**
    - Significant p-value
    - No indication of overdispersion

Logistic regression

- Testing for multicollinearity:
  - Values of 1/ vif(my_model) should be below 10

tokfreq	maxfreq	avfreq	typfreq
0.8429030	0.1652307	0.1682047	0.8062655

Logistic regression

- Testing for multicollinearity:
  - Values of 1/vif(my_model) should be below 10

tokfreq	maxfreq	avfreq	typfreq
0.8429030	0.1652307	0.1682047	0.8062655

- Testing for linearity of the logit:
  - Create interaction terms for each of the variables with its log
  - Add these to the model
  - Interaction variables should not be significant

Logistic regression

- Testing for multicollinearity:
  - Values of 1/ vif(my_model) should be below 10

tokfreq	maxfreq	avfreq	typfreq
0.8429030	0.1652307	0.1682047	0.8062655

- Testing for linearity of the logit:
  - Create interaction terms for each of the variables with its log
  - Add these to the model
  - Interaction variables should not be significant

OEVglm2 <- glm(status~ tokfreq + maxfreq + avfreq + typfreq + logtokInt + logmaxInt + logavInt + logtypInt, data=OEV, family=binomial)

LogTokInt	0.0453
LogMaxInt	0.4401
LogAvInt	0.5243
LogTypInt	0.3100

Logistic regression

- Final model
- Based on token frequency and type frequency

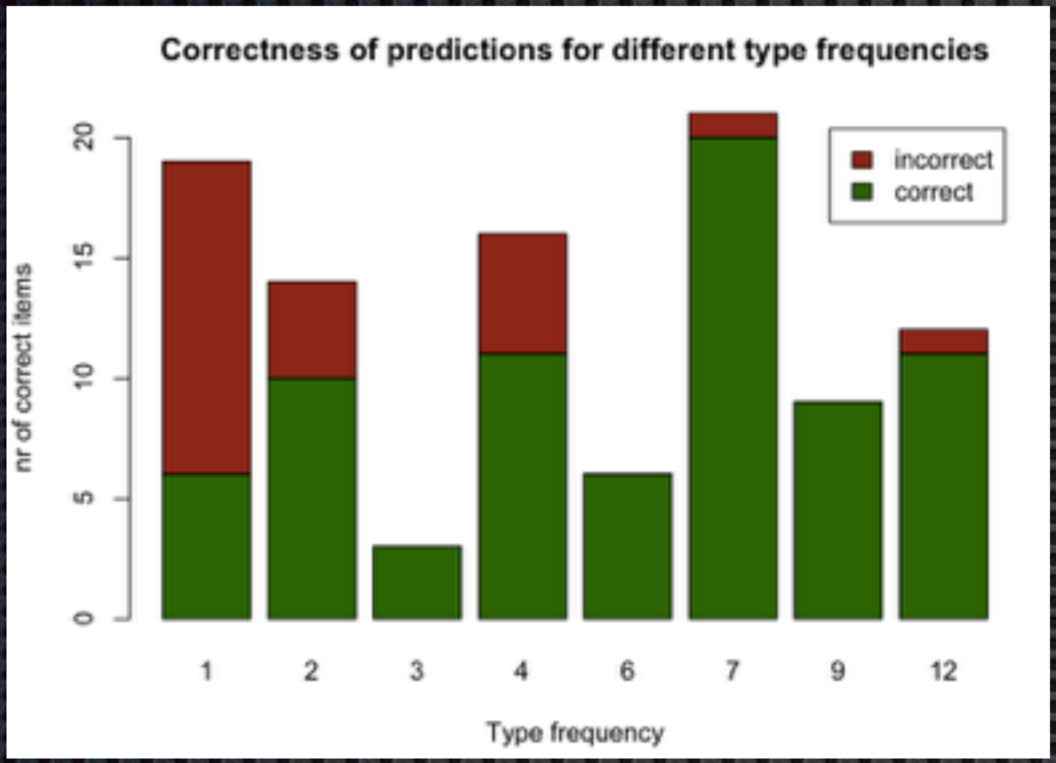
(Intercept)	-2.225e+00	5.698e-01	-3.906	9.39e-05	***
tokfreq	1.907e-05	5.842e-06	3.264	0.001098	**
typfreq	2.959e-01	7.947e-02	3.723	0.000197	***

Null deviance: 135.203 on 97 degrees of freedom
Residual deviance: 99.253 on 95 degrees of freedom
AIC: 105.25

- Chi-square = 35.94977, $p < 0.001$

Evaluation of machine

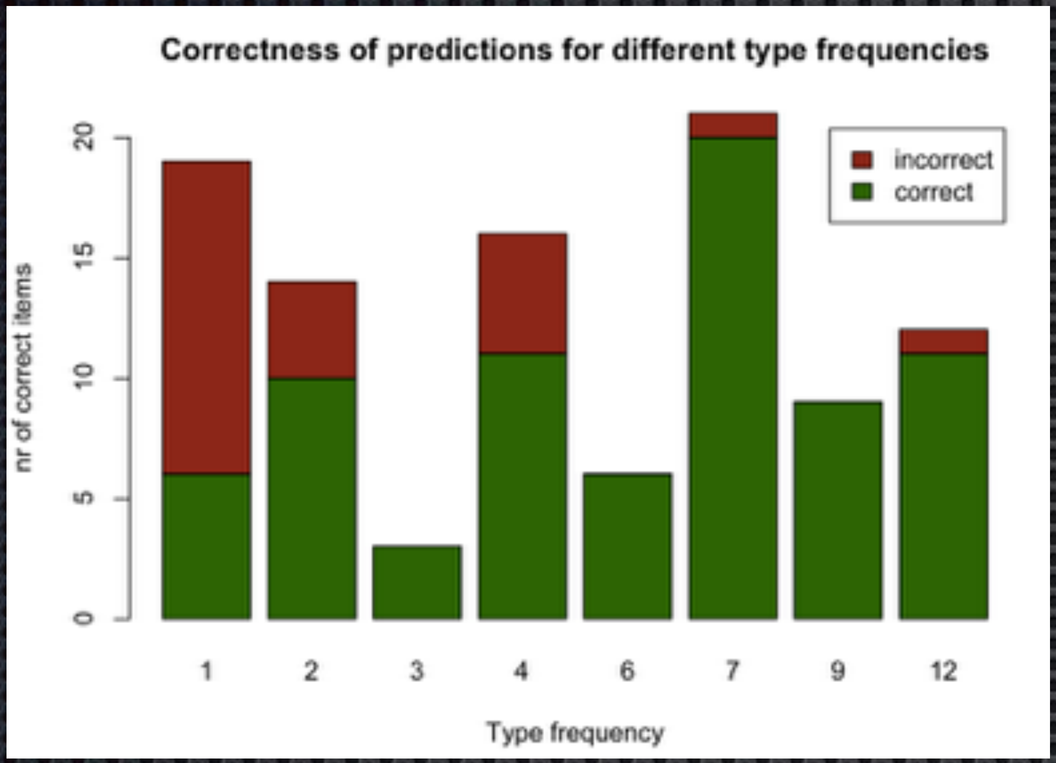
Tentative conclusions:

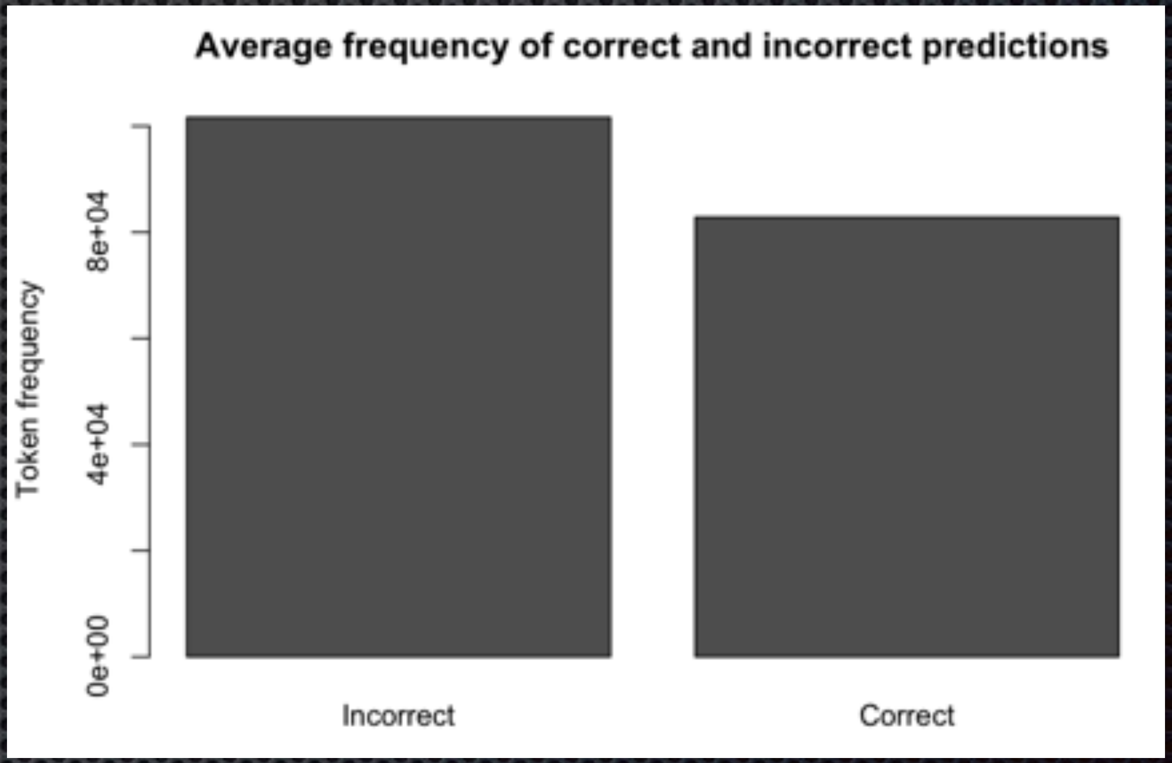


1) Small classes are "weaker"

Evaluation of machine

Tentative conclusions:





1) Small classes are “weaker”

2) Infrequent forms are “weaker”

Discussion

- Research questions were:

1. Is the stability of English strong verbs influenced by the average frequency of its analogically related forms?

2. Is the stability of English strong verbs influenced by the maximally frequent form of its analogically related forms?

- Was my methodology appropriate for answering these questions?

Discussion

- Research questions were:

1. Is the stability of English strong verbs influenced by the average frequency of its analogically related forms?

2. Is the stability of English strong verbs influenced by the maximally frequent form of its analogically related forms?

- Was my methodology appropriate for answering these questions?

    - Validity of concepts
    - Reliability
    - Validity of statistical analysis
    - Other issues

Discussion

- Validity of concepts
- Reliability
- Validity of statistical analysis
- Other issues

Discussion

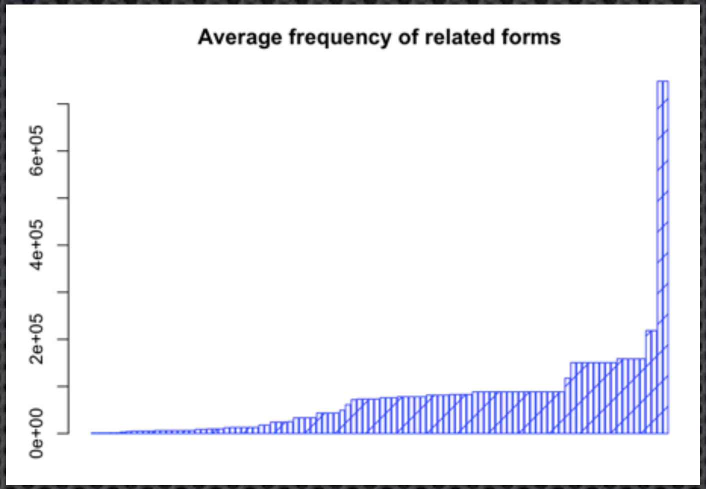
- Validity of concepts
  - Problem of collinearity between form frequency and the frequency of the class
  - Problem of testing influence on highly frequent forms when we are really only expecting related-form-frequency to matter for infrequent-yet-stable verbs
- Reliability
- Validity of statistical analysis
- Other (technical) issues

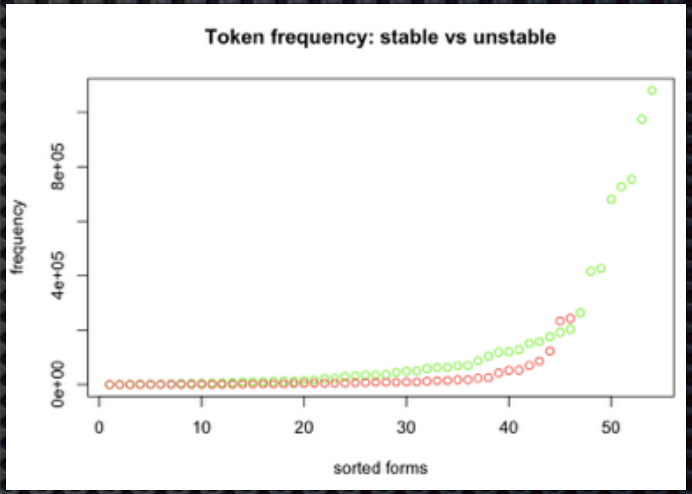
Discussion

- ~~Validity of concepts~~
- ~~Reliability~~
- **Validity of statistical analysis**
- Other (technical) issues

Discussion

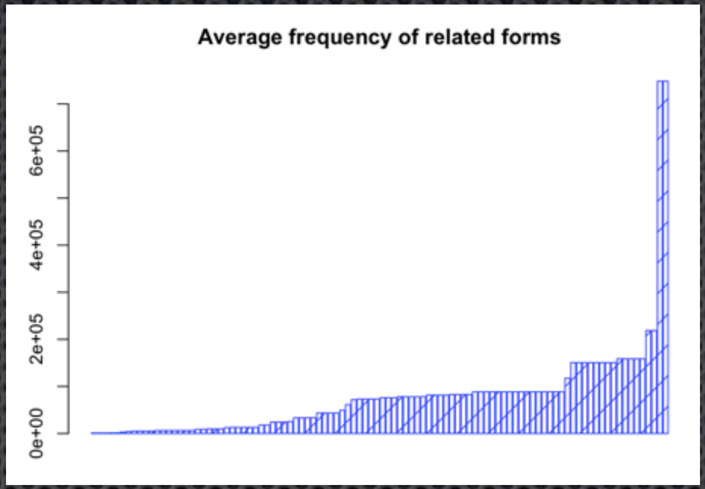
- Validity of statistical analysis
  - linearity with frequency data?

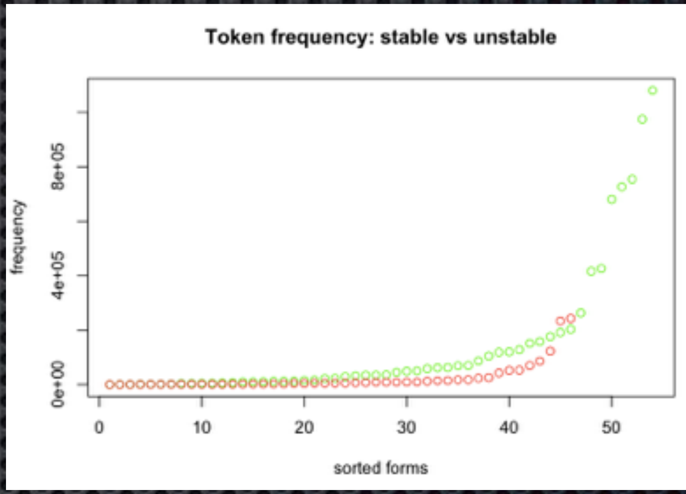




Discussion

- Validity of statistical analysis
    - linearity with frequency data?



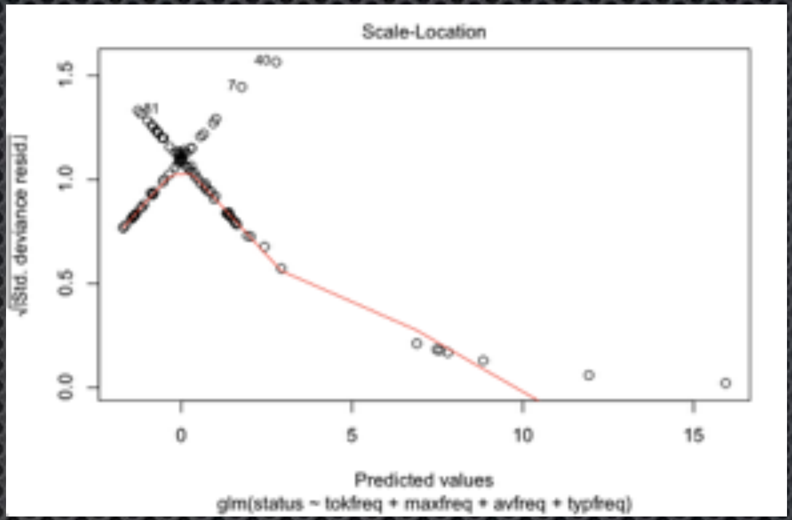


"Whilst [logistic regression] does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds. Otherwise the test underestimates the strength of the relationship and rejects the relationship too easily, that is being not significant (not rejecting the null hypothesis) where it should be significant. A solution to this problem is the categorization of the independent variables. That is transforming metric variables to ordinal level and then including them in the model.

Excerpt from: http://www.statisticssolutions.com/assumptions-of-logistic-regression/

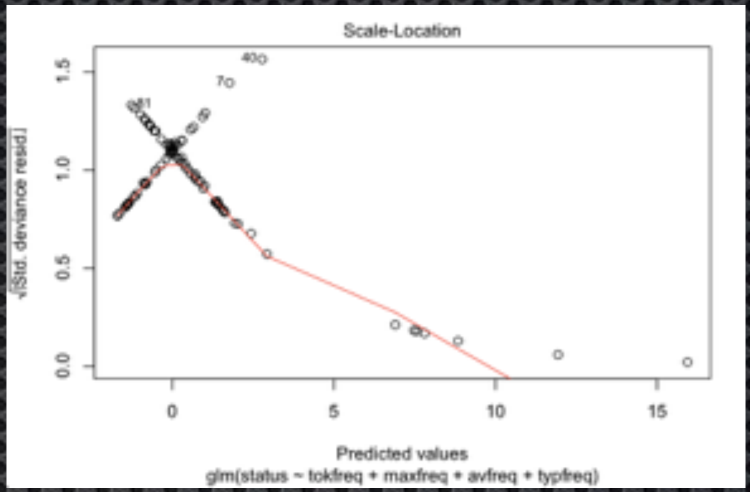
Discussion

- Other issues
    - visualizing residuals?



Discussion

- Other issues
  - visualizing residuals?



  - comparing models using anova?
    - Used in Baayen Ch 6
    - But discussion among users of R seems to suggest that the meaningfulness of such comparisons is highly debatable

NESUG 2007 **Statistics and Data Analysis**

Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use

Peter L. Flom, National Development and Research Institutes, New York, NY

David L. Cassell, Design Pathways, Corvallis, OR

Temporary conclusion

- Null hypotheses were:

1. The stability of English strong verbs is not influenced by the average frequency of its analogically related forms
2. The stability of English strong verbs is not influenced by the maximally frequent form of its analogically related forms?

→ These hypotheses may not be rejected based on a log linear regression model which treats independent variables as independent, continous variables

- Only token frequency and type frequency significantly affected stability ($\beta = 1.23$, $p < .005$ and $\beta = 2.76$, $p < .001$), confirming findings in previous studies

Temporary conclusion

- Null hypotheses were:

1. The stability of English strong verbs is not influenced by the average frequency of its analogically related forms

2. The stability of English strong verbs is not influenced by the maximally frequent form of its analogically related forms?

→ These hypotheses may not be rejected based on a log linear regression model which treats independent variables as independent, continous variables

- Only token frequency and type frequency significantly affected stability (β = 1.23, $p < .005$ and β = 2.76, $p < .001$), confirming findings in previous studies

Next: Transforming independent variables into ordinal data and performing new analyses

References

- Abbott, O. L. (1957). The Preterit and Past Participle of Strong Verbs in Seventeenth-Century American English. American Speech, 31-42.
- Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. Cognition, 90(2), 119-161.
- Albright, A., & Hayes, B. (2002, July). Modeling English past tense intuitions with minimal generalization. In Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6 (pp. 58-69). Association for Computational Linguistics.
- Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using R. Cambridge University Press.
- Crawley, M. J. (2014). Statistics: an introduction using R. John Wiley & Sons.
- Dorscheidt, T., Valchev, N., van Zoelen, T., & Nerbonne, J. Minimal generalization of Dutch diminutives.
- Field, A. P., Miles, J., & Field, Z. (2012). Discovering statistics using R. London: Sage.
- Krygier, M. (1994). The disintegration of the English strong verb system. Lang.

Questions?