A Logistic Regression model of the changing English preterit

Research results - Methodology & Statistics for Linguistic 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

<table>
<thead>
<tr>
<th></th>
<th>present</th>
<th>praeterit</th>
</tr>
</thead>
<tbody>
<tr>
<td>grow</td>
<td>grew</td>
<td>grow</td>
</tr>
<tr>
<td>claw</td>
<td>clew</td>
<td>clawed</td>
</tr>
<tr>
<td>saw</td>
<td>sawed</td>
<td>sawed</td>
</tr>
</tbody>
</table>
Introduction

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

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

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.
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.

<table>
<thead>
<tr>
<th>present</th>
<th>past tense</th>
</tr>
</thead>
<tbody>
<tr>
<td>drive</td>
<td>drove</td>
</tr>
<tr>
<td>ride</td>
<td>rode</td>
</tr>
<tr>
<td>strive</td>
<td>strove</td>
</tr>
<tr>
<td>dive</td>
<td>dove</td>
</tr>
</tbody>
</table>
Analogical Modeling

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: continuous
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: continuous

-> Logistic regression
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

\[ \text{logit}(p) = \log \left( \frac{p}{1 - p} \right) \]
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

\[ \text{logit}(p) = \log \left( \frac{p}{1 - p} \right) \]
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

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

status (stable or changed)
## Variables

- **token frequency** - [www.wordandphrase.info/](http://www.wordandphrase.info/)
- average frequency
- maximum frequency
- type frequency

<table>
<thead>
<tr>
<th>form</th>
<th>correct</th>
<th>ranking</th>
<th>missing</th>
<th>confidence</th>
<th>token freq</th>
<th>average freq</th>
<th>max freq</th>
<th>type freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1~100</td>
<td>TRUE or FALSE</td>
<td>1~3</td>
<td>TRUE or FALSE</td>
<td>0~1</td>
<td>corpus counts</td>
<td>~80 000</td>
<td>~200 000</td>
<td>1~12</td>
</tr>
</tbody>
</table>

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

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

- status (stable or changed)

<table>
<thead>
<tr>
<th>form</th>
<th>correct</th>
<th>ranking</th>
<th>missing</th>
<th>confidence</th>
<th>token freq</th>
<th>average freq</th>
<th>max freq</th>
<th>type freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1~100</td>
<td>TRUE or FALSE</td>
<td>1~3</td>
<td>TRUE or FALSE</td>
<td>0~1</td>
<td>corpus counts</td>
<td>~80 000</td>
<td>~200000</td>
<td>1~12</td>
</tr>
</tbody>
</table>
Variables

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

<table>
<thead>
<tr>
<th>form</th>
<th>correct</th>
<th>ranking</th>
<th>missing</th>
<th>confidence</th>
<th>token freq</th>
<th>average freq</th>
<th>max freq</th>
<th>type freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1~100</td>
<td>TRUE or FALSE</td>
<td>1~3</td>
<td>TRUE or FALSE</td>
<td>0~1</td>
<td>corpus counts</td>
<td>~80 000</td>
<td>~200 000</td>
<td>1~12</td>
</tr>
</tbody>
</table>

status (stable or changed)
Method

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

- token frequency - [www.wordandphrase.info/](http://www.wordandphrase.info/)
- average frequency
- maximum frequency
- type frequency

<table>
<thead>
<tr>
<th>form</th>
<th>pattern</th>
<th>form1</th>
<th>form2</th>
<th>A</th>
<th>B</th>
<th>Chang</th>
<th>Pres</th>
<th>P</th>
<th>scope</th>
<th>hits</th>
<th>reliability</th>
<th>confidence</th>
<th>related forms</th>
<th>exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>14</td>
<td>b1nd</td>
<td>b2nd</td>
<td></td>
<td></td>
<td>1/2</td>
<td>/</td>
<td>X</td>
<td></td>
<td></td>
<td>0.75</td>
<td>0.525862123591678</td>
<td>b1nd, f1nd, w1nd</td>
<td>b1nd</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>b1nd</td>
<td>b1ndld</td>
<td></td>
<td></td>
<td></td>
<td>/</td>
<td>X</td>
<td></td>
<td></td>
<td>0.4168149949</td>
<td>0.46666666666666666</td>
<td>b1nd</td>
<td>b1nd, b1ndld, blld, bld, blld, bld, b1nd, blld, b1ndld</td>
</tr>
</tbody>
</table>

- b1nd, b1ndld, blld, bld, blld, bld, b1nd, blld, b1ndld
## Method

- **token frequency** - [www.wordandphrase.info/](http://www.wordandphrase.info/)
- average frequency
- maximum frequency
- type frequency

<table>
<thead>
<tr>
<th>form</th>
<th>pattern</th>
<th>form1</th>
<th>form2</th>
<th>A</th>
<th>B</th>
<th>Change</th>
<th>Pres</th>
<th>P</th>
<th>scope</th>
<th>hits</th>
<th>reliability</th>
<th>confidence</th>
<th>related forms</th>
<th>exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>14</td>
<td>b1nd</td>
<td>b2nd</td>
<td>1 &gt;</td>
<td>2</td>
<td>1/2/2</td>
<td>/</td>
<td>X</td>
<td>4</td>
<td>3</td>
<td>0.75</td>
<td>0.525862123591678</td>
<td>b1nd, f1nd, w1nd</td>
<td>b1nd</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>b1nd</td>
<td>b1ndld</td>
<td>0 -</td>
<td>45</td>
<td>-</td>
<td>/</td>
<td>d</td>
<td>45</td>
<td>21</td>
<td>0.4666666666666666</td>
<td>0.416814099949</td>
<td>ad, av3d, b1nd, bOrd</td>
<td>b1nd, bEnd, bld, bld, blld, blld, blld, db1nd</td>
</tr>
</tbody>
</table>
Method

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

<table>
<thead>
<tr>
<th>form</th>
<th>pattern</th>
<th>Form1</th>
<th>form2</th>
<th>A</th>
<th>B</th>
<th>Change</th>
<th>Pres</th>
<th>scop</th>
<th>hits</th>
<th>reliability</th>
<th>confidence</th>
<th>related forms</th>
<th>exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>14b1nd</td>
<td>-</td>
<td>b2nd</td>
<td>1</td>
<td></td>
<td>1/2/2</td>
<td>/</td>
<td>X</td>
<td>4</td>
<td>0.75</td>
<td>0.525862123591678</td>
<td>b1nd, f1nd, w1nd</td>
<td>b1nd</td>
</tr>
<tr>
<td>14</td>
<td>4b1nd</td>
<td>-</td>
<td>b1ndd</td>
<td>2</td>
<td></td>
<td>1/2/2</td>
<td>/</td>
<td>X</td>
<td>45</td>
<td>0.4666666666666666</td>
<td>0.416814923149</td>
<td>ad, av3d, b1nd, bRand</td>
<td>b1nd, bRand, bld, blid, bisd, db1d</td>
</tr>
</tbody>
</table>
Method

- token frequency - [www.wordandphrase.info](http://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
First impression of Data

Phonological Learner File
Adam Albright/Bruce Hayes
English
Pretend Languages
Morphological categories:
Present Past
Input forms:
ls 1st
ad adId
adjust adjustId
admIt admItId
adrEs adrEst
aksEpt aksEptId
aksEs aksEst|
akt aktId
aNgyr aNgyrId
ansyr ansyrId
asUm asUmId
av3d av3dId
b1 b0t
b1nd b2nd
b1t bIt
bat batId
batyl batylId
bek bekt
bEnd bEnt
bEr bor
bes best
First impression of Data

<table>
<thead>
<tr>
<th>form</th>
<th>pattern</th>
<th>form1</th>
<th>form2</th>
<th>A</th>
<th>B</th>
<th>Change</th>
<th>Pres</th>
<th>P</th>
<th>scope</th>
<th>hits</th>
<th>reliability</th>
<th>confidence</th>
<th>related forms</th>
<th>exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>14</td>
<td>b1nd</td>
<td>b2nd</td>
<td>1</td>
<td>2</td>
<td>1/2/2</td>
<td>/</td>
<td>X</td>
<td>4</td>
<td>3</td>
<td>0.75</td>
<td>0.52586212359</td>
<td>b1nd, f1nd, w1nd</td>
<td>b1nd</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>b1nd</td>
<td>b1nd/d</td>
<td>1</td>
<td>4</td>
<td>4/0</td>
<td>/</td>
<td>d</td>
<td>45</td>
<td>21</td>
<td>0.4666666666666666</td>
<td>0.41681499949</td>
<td>ad, ay/d, b1nd, b0nd, b1nd, d1nd</td>
<td>b1nd, bEnd, blb, bld, bld, bld, dlv/1d</td>
</tr>
</tbody>
</table>

Average frequency of related forms

Maximum frequency of related forms
First impression of Data
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   -3.1900       0.001423 **
tokfreq            1.232e-05   4.782e-06     2.5770       0.009971 **
maxfreq            -1.821e-06   2.011e-06   -0.9050       0.365309
avfreq             8.086e-06   1.124e-05     0.7190       0.472000
typfreq            2.763e-01   8.030e-02     3.4400       0.000581 ***

---
Signif. codes:  *** 0.001 *** 0.01 *** 0.05 ' ' 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   -3.1906 0.001423 **
tokfreq     1.232e-05  4.782e-06    2.5700 0.009971 **
maxfreq     -1.821e-06  2.011e-06   -0.9019 0.365309
avfreq      8.086e-06  1.124e-05    0.7199 0.472000
typfreq     2.763e-01  8.030e-02   3.4460 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: 118.11

Number of Fisher Scoring iterations: 6
```
Logistic regression

Exponentiated coefficients:

<table>
<thead>
<tr>
<th>Intercept</th>
<th>tokfreq</th>
<th>maxfreq</th>
<th>avfreq</th>
<th>typfreq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1348249</td>
<td>1.0000123</td>
<td>0.9999982</td>
<td>1.0000081</td>
<td>1.3182270</td>
</tr>
</tbody>
</table>
Logistic regression

Exponentiated coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>tokfreq</th>
<th>maxfreq</th>
<th>avfreq</th>
<th>typfreq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp</td>
<td>0.1348249</td>
<td>1.00000</td>
<td>0.99999</td>
<td>1.00000</td>
<td>1.31823</td>
</tr>
</tbody>
</table>

2.5 %  97.5 %

<table>
<thead>
<tr>
<th></th>
<th>token freq</th>
<th>max freq</th>
<th>aver. freq</th>
<th>type freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp</td>
<td>1.0000022</td>
<td>1.000017</td>
<td>1.000017</td>
<td>1.5642889</td>
</tr>
</tbody>
</table>

Deviance Residuals:

-2.38296 -0.85382 0.03042 0.89985 1.74047

Exponentiated coefficients:

- token freq: 1.00000223 1.0000229
- max freq: 0.99999404 1.0000017
- average freq: 0.9999913 1.0000017
- type freq: 1.13813048 1.5642889

Number of Fisher Scoring iterations: 6
Logistic regression

- Multiple logistic regression 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 unrelated forms do not contribute to a stable outcome.

### Exponentiated coefficients:

<table>
<thead>
<tr>
<th>Intercept</th>
<th>tokfreq</th>
<th>maxfreq</th>
<th>avfreq</th>
<th>typfreq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1348249</td>
<td>1.0000123</td>
<td>0.9999982</td>
<td>1.0000081</td>
<td>1.3182270</td>
</tr>
</tbody>
</table>
Logistic regression

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.

Exponentiated coefficients:

<table>
<thead>
<tr>
<th>Intercept</th>
<th>tokfreq</th>
<th>maxfreq</th>
<th>avfreq</th>
<th>typfreq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1348249</td>
<td>1.0000123</td>
<td>0.999982</td>
<td>1.000081</td>
<td>1.318270</td>
</tr>
</tbody>
</table>
Logistic regression

```
glm(status ~ tokfreq + maxfreq + avfreq + typfreq)
```
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)

- \[ 1 - \text{pchisq} \text{difference_in_deviance, difference_in_df} \rightarrow 0.0000052948 \]
  - Significant p-value
  - No indication of overdispersion
Logistic regression

- Testing for multicollinearity:
  - Values of $1 / \text{vif(my_model)}$ should be below 10

<table>
<thead>
<tr>
<th>tokfreq</th>
<th>maxfreq</th>
<th>avfreq</th>
<th>typfreq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8429030</td>
<td>0.1652307</td>
<td>0.1682047</td>
<td>0.8062655</td>
</tr>
</tbody>
</table>
Logistic regression

- Testing for multicollinearity:
  - Values of $1 / vif(my\_model)$ should be below 10

- 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

<table>
<thead>
<tr>
<th>tokfreq</th>
<th>maxfreq</th>
<th>avfreq</th>
<th>typfreq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8429030</td>
<td>0.1652307</td>
<td>0.1682047</td>
<td>0.8062655</td>
</tr>
</tbody>
</table>
Logistic regression

- Testing for multicollinearity:
  - Values of $1/\text{vif(}\text{my}\_\text{model})$ should be below 10

- 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)
```

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>tokfreq</td>
<td>maxfreq</td>
<td>avfreq</td>
<td>typfreq</td>
<td></td>
</tr>
<tr>
<td>0.8429030</td>
<td>0.1652307</td>
<td>0.1682047</td>
<td>0.8062655</td>
<td></td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LogTokInt</td>
<td>0.0453</td>
</tr>
<tr>
<td>LogMaxInt</td>
<td>0.4401</td>
</tr>
<tr>
<td>LogAvInt</td>
<td>0.5243</td>
</tr>
<tr>
<td>LogTypInt</td>
<td>0.3100</td>
</tr>
</tbody>
</table>
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

**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?

- 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.

These hypotheses may not be rejected based on a log linear regression model which treats independent variables as independent, continuous variables.
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?

- 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.

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

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


Questions?