Unit #8: The frequency of passives

Stefan Evert

28 November 2015

Preliminaries

```
library(SIGIL)
library(effects)
library(lattice)
```

In this exercise, we will try to answer the question whether there is a significant difference between the frequency of passives in American English and in British English. While this has repeatedly been claimed in the literature, these analyses are based on an invalid application of tests for contingency tables to pooled frequency counts. Here, we will use a more appropriate linear regression model in order to take differences between individual texts – and the resulting smaller effective sample size – into account.

Note that we use a standard linear model (LM) instead of the more appropriate binomial generalized linear model (GLM) for reasons of simplicity. You can find GLM example code in Unit #8 of the SIGIL course.

The SIGIL package includes a data frame with per-text frequency counts for passive and active VPs in the extended Brown Family of corpora (see `?PassiveBrownFam`).

```
table(PassiveBrownFam$corpus)
```

```
##
##  BLOB Brown   LOB Frown  FLOB
##   500   500   500   499   500
```

Let us first select the four corpora analysed in the literature, so we can compare AmE vs. BrE in the 1960s vs. 1990s.

```
BF <- subset(PassiveBrownFam, corpus != "BLOB")
```

Note that the `corpus` variable is a so-called "factor" and remembers there all three categories ("levels" of the factor) even though `BF` no longer contains any texts from the 1930s.

```
table(BF$period)
```

```
##
## 1930 1960 1990
##    0 1000  999
```

```
BF <- droplevels(BF) # remove unused factor levels
table(BF$period)
```

```
##
## 1960 1990
## 1000  999
```

Linear models based on metadata

The goal of the linear model is to predict the relative frequency of passives (`p.pass`) based on various factors such as language variety (AmE/BrE), time period (1960/1990) or text genre using an equation of the form

$$p_i = \beta_0 + \beta_{\text{AmE/BrE}} + \beta_{1960/1990} + \beta_{\text{genre}} + \ldots + \epsilon_i.$$

The parameters β will be chosen so as to minimize the error sum of squares (ESS)

$$\text{ESS} = \sum_{i=1}^{n} \epsilon_i^2.$$

The goodness-of-fit of a "trained" LM is measure by the relative reduction in ESS compared to the baseline model $p_i = \beta_0 + \epsilon_i$, which corresponds to the variance of the dependent variable p_i. For this reason, we can think of the goodness-of-fit measure R^2 as the percentage of variance "explained" by the LM.

Let us fit a first model that only considers differences between the language varieties and time periods:

```
lm1 <- lm(p.pass ~ lang + period, data=BF)
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: p.pass
##             Df Sum Sq Mean Sq F value    Pr(>F)
## lang         1    958  957.61  12.053  0.000528 ***
## period       1   1845 1844.79  23.221 1.553e-06 ***
## Residuals 1996 158575   79.45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of variance consecutively tests each factor for significance, i.e. whether it explains significantly more variance than the previous factors alone. In this case, both language variety and time period are highly significant. A summary of the model shows the effect sizes with standard errors in a rather unintuitive form:

```
summary(lm1)
```

```
##
## Call:
## lm(formula = p.pass ~ lang + period, data = BF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.649  -6.220  -1.754   3.755  53.867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.8754     0.3452  40.191  < 2e-16 ***
## langBrE       1.3852     0.3987   3.474 0.000523 ***
## period1990   -1.9213     0.3987  -4.819 1.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.913 on 1996 degrees of freedom
## Multiple R-squared:  0.01737,    Adjusted R-squared:  0.01638
## F-statistic: 17.64 on 2 and 1996 DF,  p-value: 2.554e-08
```

It is slightly more intuitive to compute confidence intervals for the model parameters based on their standard errors

```r
confint(lm1)
```

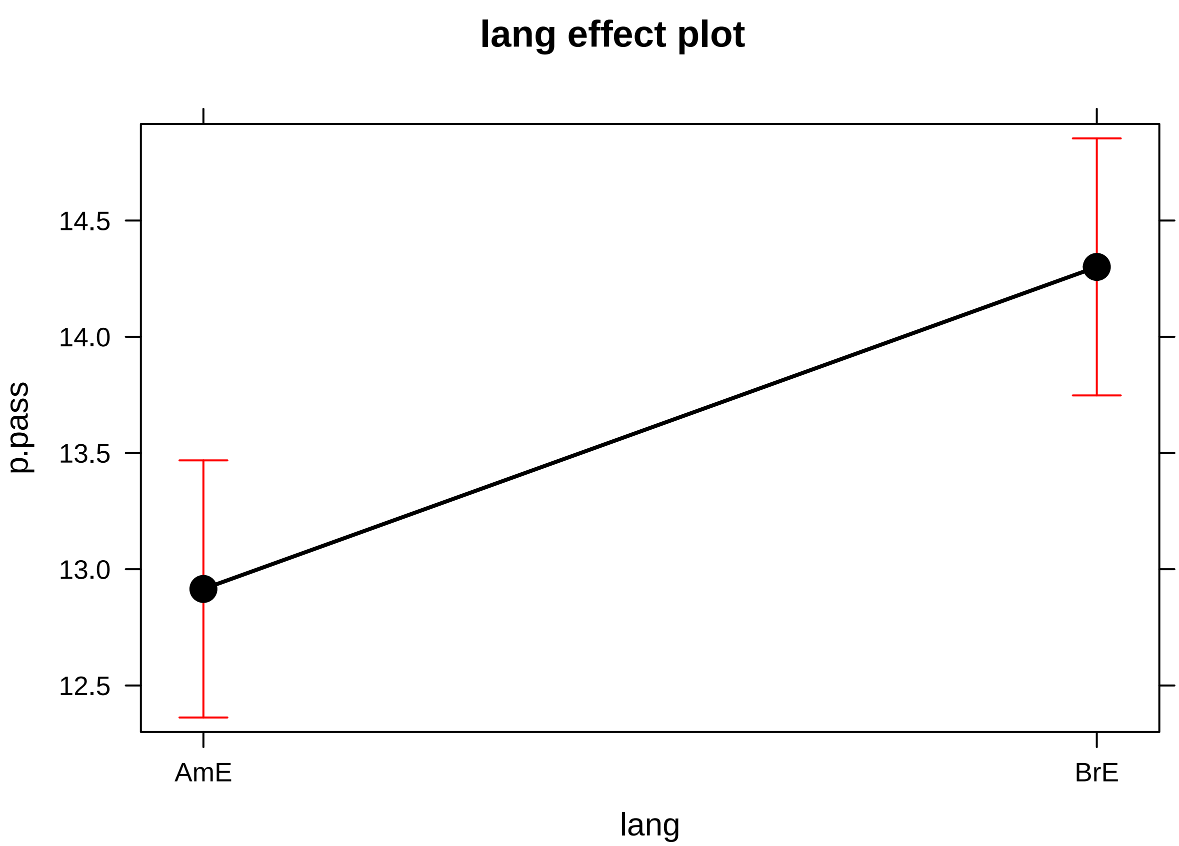
```
##                  2.5 %    97.5 %
## (Intercept) 13.1983627 14.552494
## langBrE      0.6032814  2.167158
## period1990  -2.7032449 -1.139368
```

but a much better approach is to compute and visualize the *partial effects* of each factor:

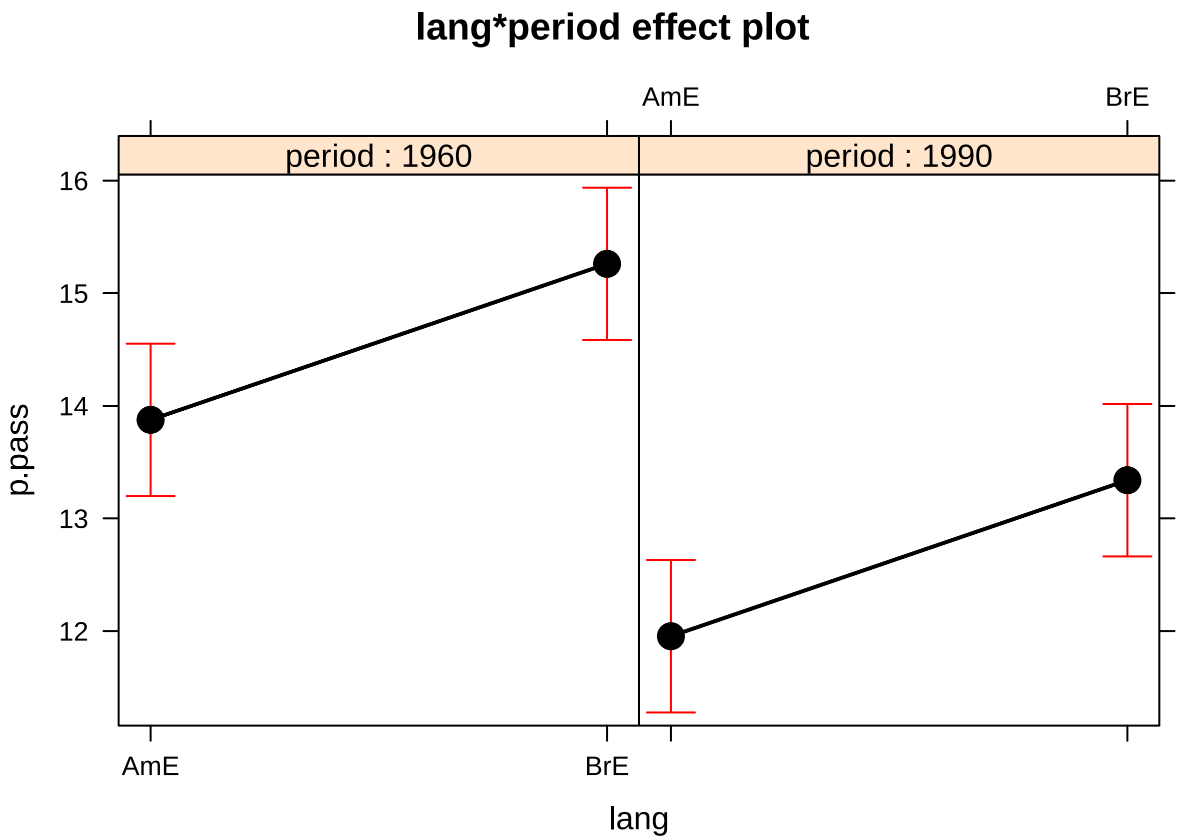
```r
Effect("lang", lm1)
```

```
##
##  lang effect
## lang
##      AmE      BrE
## 12.91526 14.30047
```

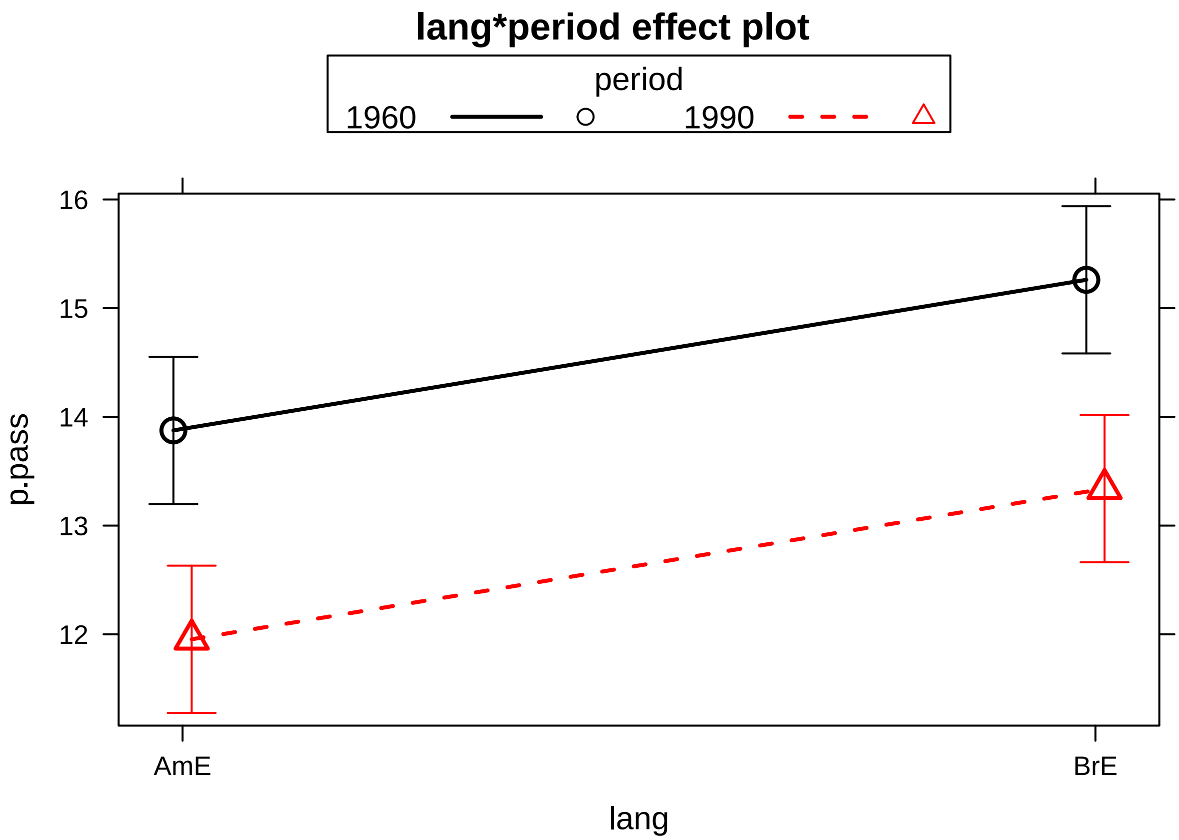
```r
plot(Effect("lang", lm1))
```




```r
plot(Effect(c("lang", "period"), lm1)) # combined effect
```




```r
plot(Effect(c("lang", "period"), lm1), multiline=TRUE, ci.style="bars")
```

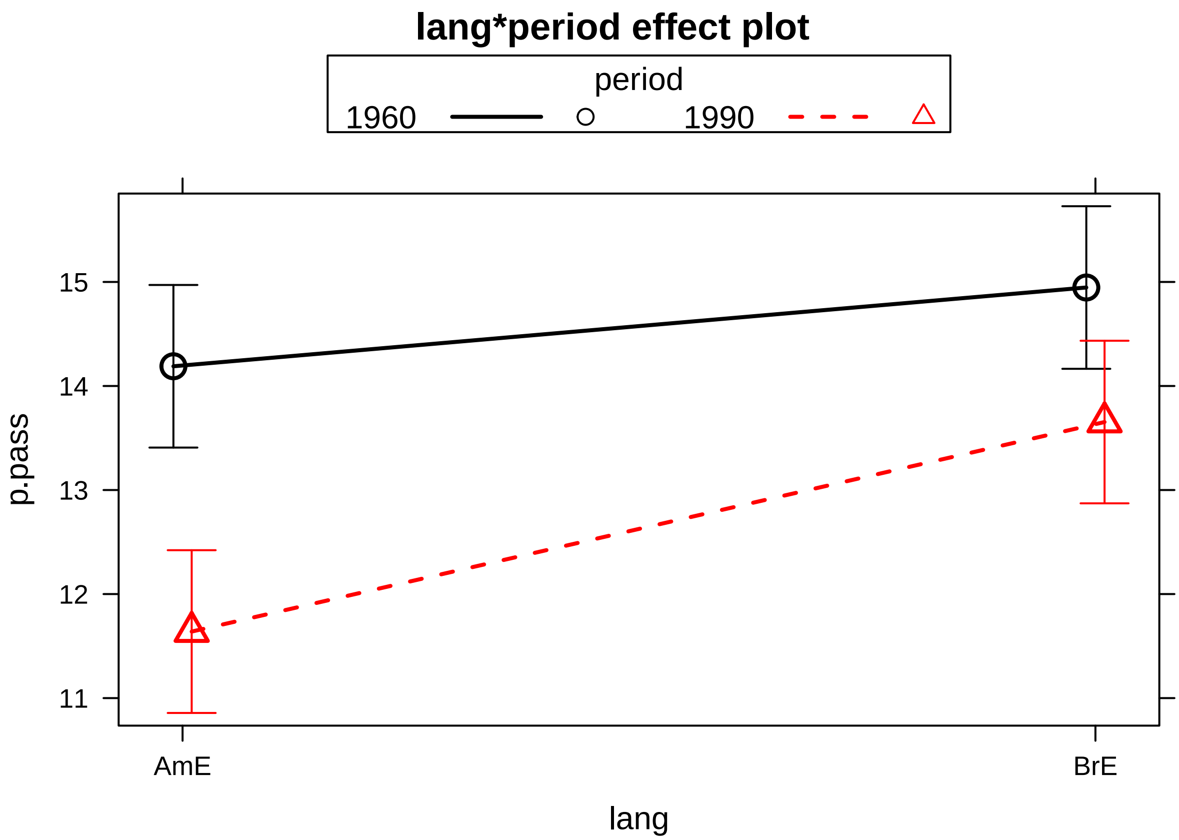


The summary above also reveals that this LM only explains 1.6% of the variance, which is highly unsatisfactory. One possible reason is that there may be an interaction between the two factors (i.e. the difference between AmE and BrE changes between the 1990s and the 1960s). Let us fit a second LM with an *interaction effect*:

```
lm2 <- lm(p.pass ~ lang * period, data=BF)
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: p.pass
##                Df Sum Sq Mean Sq F value    Pr(>F)
## lang            1    958  957.61 12.0625 0.0005254 ***
## period          1   1845 1844.79 23.2378  1.54e-06 ***
## lang:period     1    197  197.45  2.4871 0.1149394
## Residuals    1995 158378   79.39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(Effect(c("lang", "period"), lm2), multiline=TRUE, ci.style="bars")
```



While the difference is more pronounced in the 1990s than the 1960s, this interaction effect is not significant! Let us try to account better for frequency differences between texts by including the text genre as a factor:

```
lm3 <- lm(p.pass ~ lang + period + genre, data=BF)
anova(lm3)
```

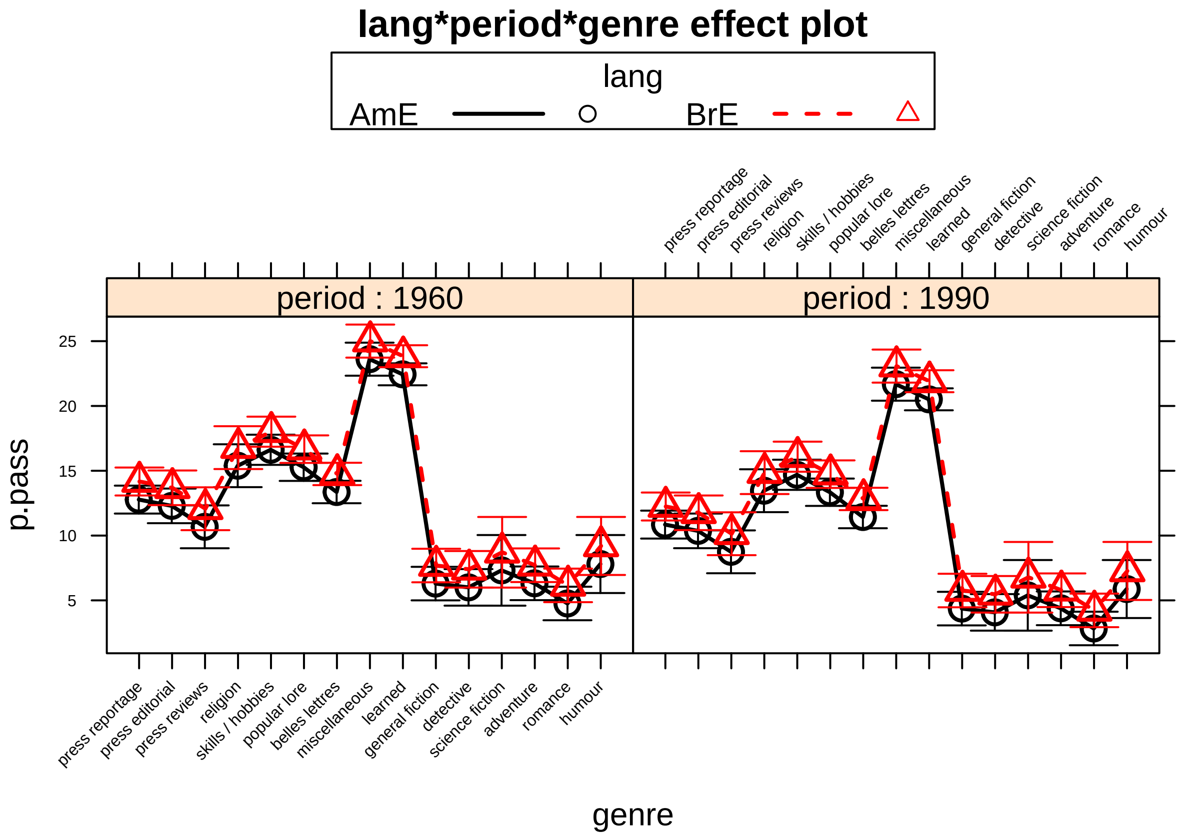
```
## Analysis of Variance Table
##
## Response: p.pass
##              Df Sum Sq Mean Sq F value    Pr(>F)
## lang          1    958   957.6  21.171 4.467e-06 ***
## period        1   1845  1844.8  40.785 2.111e-10 ***
## genre        14  68925  4923.2 108.843 < 2.2e-16 ***
## Residuals 1982  89650    45.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm3)$adj.r.squared # 44% explained variance is much better
```

```
## [1] 0.4399845
```

It's very hard to make sense of confidence intervals for the many levels of `genre`, so let us rather plot its partial effects.

```r
plot(Effect(c("lang", "period", "genre"), lm3), multiline=TRUE, ci.style="bars", rotx=45)
```



Again, there might be interactions between the three factors, so we should test their significance.

```r
lm4 <- lm(p.pass ~ lang * period * genre, data=BF)
anova(lm4)
```

```
## Analysis of Variance Table
##
## Response: p.pass
##                      Df Sum Sq Mean Sq  F value    Pr(>F)
## lang                  1    958   957.6  21.5884 3.606e-06 ***
## period                1   1845  1844.8  41.5891 1.418e-10 ***
## genre                14  68925  4923.2 110.9894 < 2.2e-16 ***
## lang:period           1    202   202.2   4.5592   0.03287 *
## lang:genre           14    601    42.9   0.9673   0.48478
## period:genre         14   2361   168.6   3.8014 2.170e-06 ***
## lang:period:genre    14    477    34.1   0.7689   0.70416
## Residuals          1939  86009    44.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Most interactions aren't significant, but R^2 has improved slightly to 0.4508143. This apparent improvement in goodness-of-fit is misleading, though, because the LM with interactions has many more parameters than the previous one, allowing it to fit random patterns in the data set. One way of assessing whether there is an actual improvement is Akaike's Information Criterion (AIC), which adjusts R^2 for the number of model parameters:

```r
AIC(lm1, lm2, lm3, lm4)
```

```
##     df      AIC
## lm1  4 14423.71
## lm2  5 14423.21
## lm3 18 13311.65
## lm4 61 13314.77
```

The AIC for `lm4` is actually worse than for `lm3`, showing that we are indeed overfitting random patterns with the interaction model. However, you may also have noticed that the interaction between language variety became highly significant in `lm4` – it is quite typical for such effects to become visible only when other sources of variation are taken into account. Let us try another model that includes only this interaction effect:

```r
lm5 <- lm(p.pass ~ lang * period + genre, data=BF)
anova(lm5)
```

```
## Analysis of Variance Table
##
## Response: p.pass
##               Df Sum Sq Mean Sq  F value    Pr(>F)
## lang           1    958   957.6  21.2081 4.382e-06 ***
## period         1   1845  1844.8  40.8564 2.037e-10 ***
## genre         14  68925  4923.2 109.0341 < 2.2e-16 ***
## lang:period    1    202   202.2   4.4788   0.03444 *
## Residuals   1981  89448    45.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Take a look at the partial effects of `lang` and `period` and their confidence intervals. Use `AIC` to confirm that this model is actually better than `lm3`. What is your (linguistic) interpretation of the analysis?

Linear models based on distributional features

As explained in the lecture slides, 44% of explained variance is still somewhat unsatisfactory, leaving a large part of the frequency differences between texts unaccounted for. We will now try to use latent features from an unsupervised distributional analysis of the Brown Family texts as additional predictors, starting from the best model so far (`lm5`). These distributional features are included in the SIGIL package (see `?DistFeatBrownFam`).

We could use the `merge` function to append these features to the data frame `BF` (needed because there is one text missing in `BF`, so the two data frames wouldn't align), but in this case there is a much easier solution. The rows of `DistFeatBrownFam` have helpfully been labelled with text IDs, so we can directly extract the desired rows:

```r
BF <- cbind(BF, DistFeatBrownFam[BF$id, -1]) # -1 removes duplicate id column
```

Here is a linear model with all latent topic dimensions and latent registers (excluding verb tags to avoid circularity). Unfortunately, the variable names have to be spelled out, but that's what cut & paste is for.

```r
lm6 <- lm(p.pass ~ lang * period + genre
          + top1 + top2 + top3 + top4 + top5 + top6 + top7 + top8 + top9
          + reg1 + reg2 + reg3 + reg4 + reg5 + reg6 + reg7 + reg8 + reg9, data=BF)
anova(lm6)
```

```
## Analysis of Variance Table
##
## Response: p.pass
##                Df Sum Sq Mean Sq  F value    Pr(>F)    
## lang            1    958   957.6  41.9276 1.194e-10 ***
## period          1   1845  1844.8  80.7717 < 2.2e-16 ***
## genre          14  68925  4923.2 215.5567 < 2.2e-16 ***
## top1            1  11807 11806.6 516.9369 < 2.2e-16 ***
## top2            1  10924 10923.8 478.2856 < 2.2e-16 ***
## top3            1    592   591.6  25.9016 3.936e-07 ***
## top4            1   5925  5925.5 259.4389 < 2.2e-16 ***
## top5            1   2732  2732.3 119.6305 < 2.2e-16 ***
## top6            1     74    74.5   3.2601  0.071137 .  
## top7            1   1458  1457.9  63.8305 2.290e-15 ***
## top8            1   3999  3999.3 175.1059 < 2.2e-16 ***
## top9            1    718   717.6  31.4191 2.373e-08 ***
## reg1            1    940   940.3  41.1697 1.745e-10 ***
## reg2            1    741   741.3  32.4549 1.404e-08 ***
## reg3            1      8     8.3   0.3621  0.547438    
## reg4            1    289   288.9  12.6472  0.000385 ***
## reg5            1   3430  3429.6 150.1609 < 2.2e-16 ***
## reg6            1    376   376.1  16.4663 5.146e-05 ***
## reg7            1    431   430.7  18.8578 1.480e-05 ***
## reg8            1      0     0.0   0.0019  0.965458    
## reg9            1      2     2.2   0.0965  0.756078    
## lang:period     1    370   370.0  16.1983 5.921e-05 ***
## Residuals    1963  44834    22.8                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have a wild mixture of significant and non-significant factors now. A common practice is to remove all predictors that do not improve the model fit by stepwise feature selection:

```r
lm7 <- step(lm6)
anova(lm7)
```

Have you noticed that the effects for language variety and time period as well as their interaction are all highly significant now? The LM with distributional features also achieves a much better goodness-of-fit of 71.7%:

```r
summary(lm7)$adj.r.squared
```

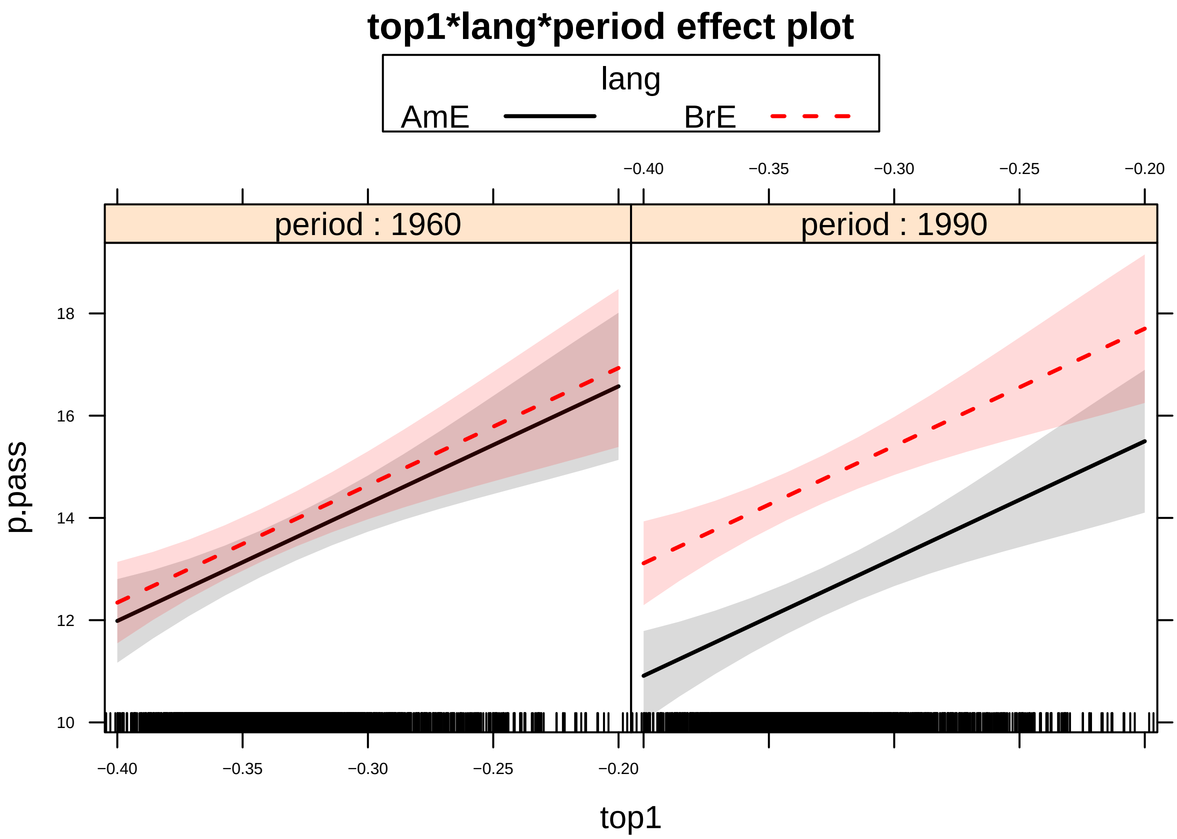
```
## [1] 0.7173145
```

```r
AIC(lm5, lm6, lm7) # stepwise selection improves AIC
```

```
##     df      AIC
## lm5 19 13309.13
## lm6 37 11964.44
## lm7 19 11946.07
```

Can you explain why the partial effects plots for the distributional features look different than before?

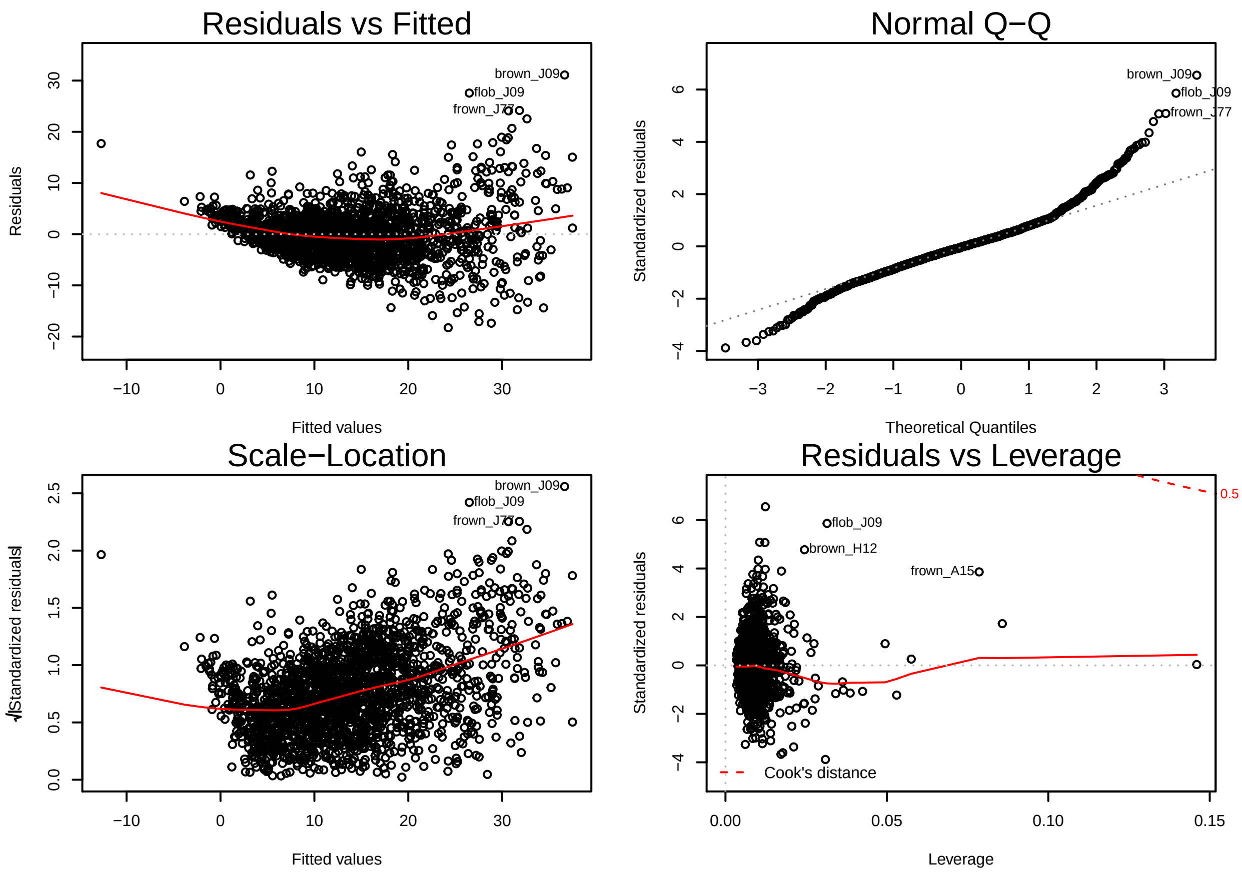
```r
plot(Effect(c("top1", "lang", "period"), lm7), multiline=TRUE, ci.style="bands")
```



If you look closely, you'll find that `genre` is no longer included in the final model as a predictive factor. Can you explain what might be going on here?

Finally, one should always look at the model diagnostics to check hints that model assumptions (such as normality or the infamous *homoscedasticity*) may be violated or that outliers may have distorted the analysis.

```r
plot(lm7)
```



Residuals vs Fitted
Residuals
Fitted values
brown_J09
flob_J09
frown_J77
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
flob_J09
brown_H12
frown_A15
Cook's distance
0.5