Categorical predictors, interactions, logistic regression, and hierarchical models

Roger Levy

Massachusetts Institute of Technology

April 25, 2024

Categorical predictors: contrast matrices

- Often we want to bring effects of categorical predictors into our studies

Categorical predictors: contrast matrices

- Often we want to bring effects of categorical predictors into our studies
  - Is the relative clause being processed subject-extracted or object-extracted (Grodner and Gibson, 2005; Traxler et al., 2002)?

    The reporter who attacked the senator (subject-extracted)
    The reporter who the senator attacked (object-extracted)

Categorical predictors: contrast matrices

- Often we want to bring effects of categorical predictors into our studies
  - Is the relative clause being processed subject-extracted or object-extracted (Grodner and Gibson, 2005; Traxler et al., 2002)?

    The reporter who attacked the senator (subject-extracted)
    The reporter who the senator attacked (object-extracted)
  - Is the vowel lax or tense?

Categorical predictors: contrast matrices

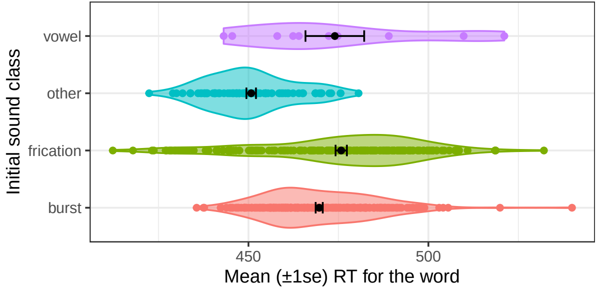
- Often we want to bring effects of categorical predictors into our studies
  - Is the relative clause being processed subject-extracted or object-extracted (Grodner and Gibson, 2005; Traxler et al., 2002)?

    The reporter who attacked the senator (subject-extracted)
    The reporter who the senator attacked (object-extracted)
  - Is the vowel lax or tense?
- Example in word naming: the category of the initial sound of the word

`burst`	*bait*
`frication`	*chess*
`other consonant`	*wrist*
`vowel`	*inch*

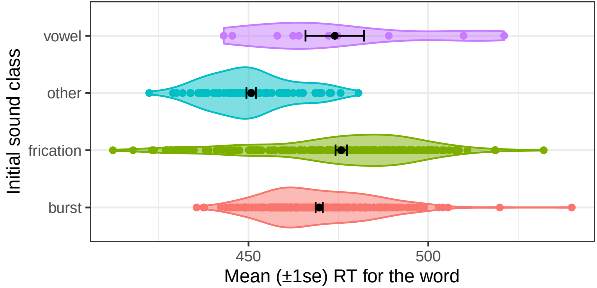
Categorical predictors

- Initial evidence suggests that there may well be some real effects of initial phone class on word naming RT:



Categorical predictors

- Initial evidence suggests that there may well be some real effects of initial phone class on word naming RT:



- **Note:** within-class variance seems to differ across classes—this is sometimes called heteroskedasticity. We won't worry about this for now, but it's something to keep in mind and we can eventually incorporate ways to deal with it

Categorical predictors

- Initial phone class is a four-predictor categorical variable

Categorical predictors

- Initial phone class is a four-predictor categorical variable
- Hence, a model which encodes a different predicted mean for each initial phone class must have four parameters (plus a residual noise parameter σ^2)

Categorical predictors

- Initial phone class is a four-predictor categorical variable
- Hence, a model which encodes a different predicted mean for each initial phone class must have four parameters (plus a residual noise parameter σ^2)
- But a model in which phone class doesn't have an effect already has one parameter (plus σ^2):

$$\mathit{RT} = \alpha + \epsilon$$

Categorical predictors

- Initial phone class is a four-predictor categorical variable
- Hence, a model which encodes a different predicted mean for each initial phone class must have four parameters (plus a residual noise parameter σ^2)
- But a model in which phone class doesn't have an effect already has one parameter (plus σ^2):

$$RT = \alpha + \epsilon$$

- To fit the categorical predictors into the model, we use a coding scheme to represent the categorical values as numeric predictors

Categorical predictors

- Initial phone class is a four-predictor categorical variable
- Hence, a model which encodes a different predicted mean for each initial phone class must have four parameters (plus a residual noise parameter σ^2)
- But a model in which phone class doesn't have an effect already has one parameter (plus σ^2):

$$RT = \alpha + \epsilon$$

- To fit the categorical predictors into the model, we use a coding scheme to represent the categorical values as numeric predictors
- Many different coding schemes available (I'll cover three); critically, these models are equivalent in the following key sense: they can express the same set of probability distributions $P(Y|X)$!

Categorical predictors

- Initial phone class is a four-predictor categorical variable
- Hence, a model which encodes a different predicted mean for each initial phone class must have four parameters (plus a residual noise parameter σ^2)
- But a model in which phone class doesn't have an effect already has one parameter (plus σ^2):

$$RT = \alpha + \epsilon$$

- To fit the categorical predictors into the model, we use a coding scheme to represent the categorical values as numeric predictors
- Many different coding schemes available (I'll cover three); critically, these models are equivalent in the following key sense: they can express the same set of probability distributions $P(Y|X)$!
- Hence, your choice of coding scheme should generally be guided by how you want to interpret model parameters

Dummy coding

With dummy coding, you drop the intercept term and create one 0/1 “dummy” predictor per value of your categorical predictor:

Level of `Init`	X_1	X_2	X_3	X_4
`burst`	1	0	0	0
`frication`	0	1	0	0
`other`	0	0	1	0
`vowel`	0	0	0	1

Dummy coding

With dummy coding, you drop the intercept term and create one 0/1 “dummy” predictor per value of your categorical predictor:

Level of `Init`	X_1	X_2	X_3	X_4
`burst`	1	0	0	0
`frication`	0	1	0	0
`other`	0	0	1	0
`vowel`	0	0	0	1

This leads to the model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

Dummy coding

With dummy coding, you drop the intercept term and create one 0/1 “dummy” predictor per value of your categorical predictor:

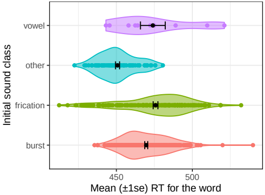
Level of `Init`	X_1	X_2	X_3	X_4
`burst`	1	0	0	0
`frication`	0	1	0	0
`other`	0	0	1	0
`vowel`	0	0	0	1

This leads to the model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

(Note: you cannot use dummy coding for more than one additively combined categorical predictor, since you can only sacrifice the intercept once)

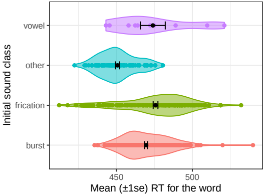
Dummy coding



Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
`burst`	469.65	1.15	$\ll 0.001$
`frication`	475.77	1.28	$\ll 0.001$
`other`	450.76	1.96	$\ll 0.001$
`vowel`	474.00	5.70	$\ll 0.001$

- Here, each parameter is just the mean RT for the corresponding initial phone type

Dummy coding



Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
`burst`	469.65	1.15	$\ll 0.001$
`frication`	475.77	1.28	$\ll 0.001$
`other`	450.76	1.96	$\ll 0.001$
`vowel`	474.00	5.70	$\ll 0.001$

- Here, each parameter is just the mean RT for the corresponding initial phone type
- The t statistics reflect the null hypothesis that mean RT for that initial phone type is 0 (not a very useful null hypothesis!)

Treatment coding

- With treatment coding, you leave the intercept in and arbitrarily label one value of the categorical predictor as the "baseline" level

Treatment coding

- With treatment coding, you leave the intercept in and arbitrarily label one value of the categorical predictor as the "baseline" level
- You then add a 0/1 predictor variable for each non-baseline level of the predictor

Treatment coding

- With treatment coding, you leave the intercept in and arbitrarily label one value of the categorical predictor as the “baseline” level
- You then add a 0/1 predictor variable for each non-baseline level of the predictor
- E.g., if we make burst the baseline level for initial phone:

Treatment coding

- With treatment coding, you leave the intercept in and arbitrarily label one value of the categorical predictor as the “baseline” level
- You then add a 0/1 predictor variable for each non-baseline level of the predictor
- E.g., if we make burst the baseline level for initial phone:

Treatment coding

- With treatment coding, you leave the intercept in and arbitrarily label one value of the categorical predictor as the "baseline" level
- You then add a 0/1 predictor variable for each non-baseline level of the predictor
- E.g., if we make burst the baseline level for initial phone:

Level of `Init`	X_1	X_2	X_3
`burst`	0	0	0
`frication`	1	0	0
`other`	0	1	0
`vowel`	0	0	1

Treatment coding

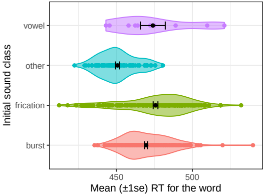
- With treatment coding, you leave the intercept in and arbitrarily label one value of the categorical predictor as the "baseline" level
- You then add a 0/1 predictor variable for each non-baseline level of the predictor
- E.g., if we make burst the baseline level for initial phone:

Level of `Init`	X_1	X_2	X_3
`burst`	0	0	0
`frication`	1	0	0
`other`	0	1	0
`vowel`	0	0	1

This leads to the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

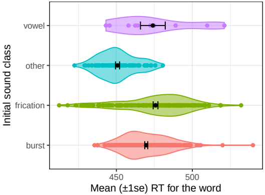
Treatment coding



Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\lvert t \rvert)$
Intercept	469.65	1.15	$\ll 0.001$
`frication`	6.12	1.72	0.000411
`other`	-18.90	2.27	$\ll 0.001$
`vowel`	4.35	5.82	0.455

- Here, weights 2–4 indicate the *difference* between the mean RT for words of the given phone type and for `burst`-initial words

Treatment coding



Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
Intercept	469.65	1.15	$\ll 0.001$
`frication`	6.12	1.72	0.000411
`other`	-18.90	2.27	$\ll 0.001$
`vowel`	4.35	5.82	0.455

- Here, weights 2–4 indicate the *difference* between the mean RT for words of the given phone type and for `burst`-initial words
- The *t*-statistic-based *p*-value indicates that fricative-initial words are slower than burst-initial words, and that words starting with other consonants are faster, but that we have insufficient information to conclude that vowel-initial words are either faster or slower

Sum (or deviation) coding

- Sum (or deviation) coding is designed to fit weights for all values of the categorical predictor, subject to the constraint that the weights sum to zero

Sum (or deviation) coding

- Sum (or deviation) coding is designed to fit weights for all values of the categorical predictor, subject to the constraint that the weights sum to zero
- The intercept remains in the model, and represents the hypothetical "average" response across all conditions

Sum (or deviation) coding

- Sum (or deviation) coding is designed to fit weights for all values of the categorical predictor, subject to the constraint that the weights sum to zero
- The intercept remains in the model, and represents the hypothetical “average” response across all conditions

Sum (or deviation) coding

- Sum (or deviation) coding is designed to fit weights for all values of the categorical predictor, subject to the constraint that the weights sum to zero
- The intercept remains in the model, and represents the hypothetical "average" response across all conditions

Level of `Init`	X_1	X_2	X_3
`burst`	1	0	0
`frication`	0	1	0
`other`	0	0	1
`vowel`	-1	-1	-1

- This leads to the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Sum (or deviation) coding

- Equivalently we can describe the sum-coding model as

Sum (or deviation) coding

- Equivalently we can describe the sum-coding model as

Sum (or deviation) coding

- Equivalently we can describe the sum-coding model as

Level of Init	X_1	X_2	X_3	X_4
burst	1	0	0	0
frication	0	1	0	0
other	0	0	1	0
vowel	0	0	0	1

- with the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

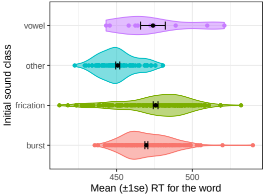
subect to the constraint that

$$\beta_1 + \beta_2 + \beta_3 + \beta_4 = 0$$

Or equivalently

$$\beta_4 = -\beta_1 - \beta_2 - \beta_3$$

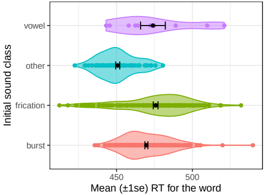
Sum (or deviation) coding



Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
Intercept	467.55	1.57	$\ll 0.001$
X_1	2.11	1.77	0.233
X_2	8.23	1.81	$\ll 0.001$
X_3	-16.79	2.09	$\ll 0.001$

- Parameters 2–4 are the *offsets* of the initial-phone-specific means from the hypothetical “grand average” (which is not the mean RT, because we have different # observations for different initial phones!)

Sum (or deviation) coding

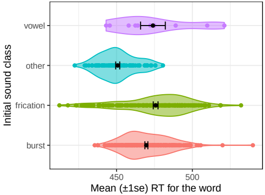


Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(> \mid t \mid)$
Intercept	467.55	1.57	$\ll 0.001$
X_1	2.11	1.77	0.233
X_2	8.23	1.81	$\ll 0.001$
X_3	-16.79	2.09	$\ll 0.001$

- Parameters 2–4 are the *offsets* of the initial-phone-specific means from the hypothetical "grand average" (which is not the mean RT, because we have different # observations for different initial phones!)
- To recover the condition-specific offset $\widehat{\beta_4}$ for and its standard error $SE(\widehat{\beta_4})$, recall that

$$\beta_4 = -\beta_1 - \beta_2 - \beta_3 \qquad \Rightarrow \qquad \widehat{\beta_4} = -\widehat{\beta_1} - \widehat{\beta_2} - \widehat{\beta_3}$$

Sum (or deviation) coding

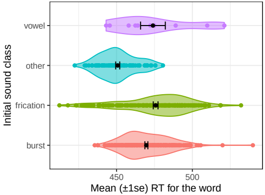


Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
Intercept	467.55	1.57	$\ll 0.001$
X_1	2.11	1.77	0.233
X_2	8.23	1.81	$\ll 0.001$
X_3	-16.79	2.09	$\ll 0.001$

- Parameters 2–4 are the *offsets* of the initial-phone-specific means from the hypothetical "grand average" (which is not the mean RT, because we have different # observations for different initial phones!)
- To recover the condition-specific offset $\widehat{\beta}_4$ for and its standard error $SE(\widehat{\beta}_4)$, recall that

$$\beta_4 = -\beta_1 - \beta_2 - \beta_3 \qquad \Rightarrow \qquad \widehat{\beta}_4 = -\widehat{\beta}_1 - \widehat{\beta}_2 - \widehat{\beta}_3$$

Sum (or deviation) coding



Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
Intercept	467.55	1.57	$\ll 0.001$
X_1	2.11	1.77	0.233
X_2	8.23	1.81	$\ll 0.001$
X_3	-16.79	2.09	$\ll 0.001$

- Parameters 2–4 are the *offsets* of the initial-phone-specific means from the hypothetical "grand average" (which is not the mean RT, because we have different # observations for different initial phones!)
- To recover the condition-specific offset $\widehat{\beta}_4$ for and its standard error $SE(\widehat{\beta}_4)$, recall that

$$\beta_4 = -\beta_1 - \beta_2 - \beta_3 \qquad \Rightarrow \qquad \widehat{\beta}_4 = -\widehat{\beta}_1 - \widehat{\beta}_2 - \widehat{\beta}_3$$

This gives us $\widehat{\beta}_4 = -6.45$; using the formula for variance of the sum of random variables we can recover that $SE(\widehat{\beta}_4) = 4.33$ and hence the p=value for that offset is 0.136

When to care about coding schema?

- To reiterate: the set of probability distributions $P(Y|X)$ that can be fitted does not depend on choice of coding schema (at least in this current type of case!)

When to care about coding schema?

- To reiterate: the set of probability distributions $P(Y|X)$ that can be fitted does not depend on choice of coding schema (at least in this current type of case!)
- So frequentist model comparisons involving the model (e..g, the F test, the likelihood ratio test) will give the same results regardless of the coding schema you choose

When to care about coding schema?

- To reiterate: the set of probability distributions $P(Y|X)$ that can be fitted does not depend on choice of coding schema (at least in this current type of case!)
- So frequentist model comparisons involving the model (e..g, the F test, the likelihood ratio test) will give the same results regardless of the coding schema you choose
  - E.g., in our case, comparing our model M_A against the “null” model $M_0 : RT = \alpha + \epsilon$ gives the result

$$\frac{(RSS_0 - RSS_A)/(m_A - m_0)}{RSS_A/(n - m_A - 1)} = \frac{(211685.57 - 173916.95)/3}{173916.95/535}$$
$$= 38.73$$

When to care about coding schema?

- To reiterate: the set of probability distributions $P(Y|X)$ that can be fitted does not depend on choice of coding schema (at least in this current type of case!)
- So frequentist model comparisons involving the model (e..g, the F test, the likelihood ratio test) will give the same results regardless of the coding schema you choose
  - E.g., in our case, comparing our model M_A against the "null" model $M_0 : RT = \alpha + \epsilon$ gives the result

$$\frac{(RSS_0 - RSS_A)/(m_A - m_0)}{RSS_A/(n - m_A - 1)} = \frac{(211685.57 - 173916.95)/3}{173916.95/535}$$
$$= 38.73$$

When to care about coding schema?

- To reiterate: the set of probability distributions $P(Y|X)$ that can be fitted does not depend on choice of coding schema (at least in this current type of case!)
- So frequentist model comparisons involving the model (e..g, the F test, the likelihood ratio test) will give the same results regardless of the coding schema you choose
  - E.g., in our case, comparing our model M_A against the "null" model $M_0 : RT = \alpha + \epsilon$ gives the result

$$\begin{aligned}\frac{(RSS_0 - RSS_A)/(m_A - m_0)}{RSS_A/(n - m_A - 1)} &= \frac{(211685.57 - 173916.95)/3}{173916.95/535} \\ &= 38.73\end{aligned}$$

  - Consulting the cumulative distribution function for $F_{3,535}$, we find that this is highly significant—$p \ll 0.001$

When to care about coding schema?

- To reiterate: the set of probability distributions $P(Y|X)$ that can be fitted does not depend on choice of coding schema (at least in this current type of case!)
- So frequentist model comparisons involving the model (e..g, the F test, the likelihood ratio test) will give the same results regardless of the coding schema you choose
  - E.g., in our case, comparing our model M_A against the "null" model $M_0 : RT = \alpha + \epsilon$ gives the result

$$\frac{(RSS_0 - RSS_A)/(m_A - m_0)}{RSS_A/(n - m_A - 1)} = \frac{(211685.57 - 173916.95)/3}{173916.95/535}$$
$$= 38.73$$

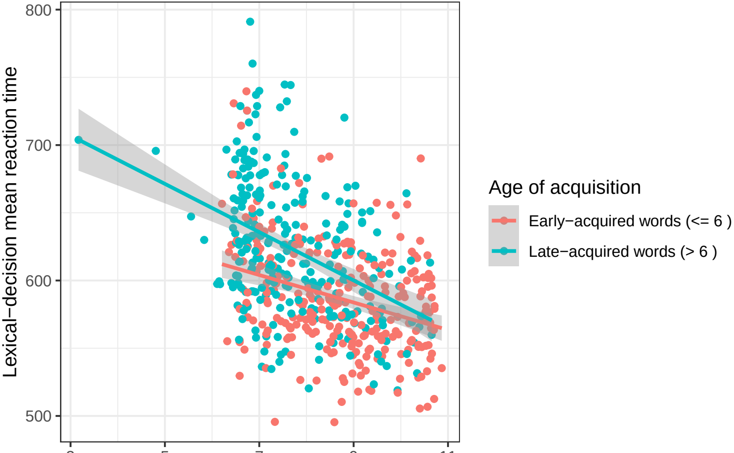
  - Consulting the cumulative distribution function for $F_{3,535}$, we find that this is highly significant—$p \ll 0.001$
  - Rather, your choice of contrasts should reflect how you want to *interpret* the parameters in your model!

Interactions between predictors

- Sometimes X_1's effect on your response Y will differ depending on the value of another predictor X_2

Interactions between predictors

- Sometimes X_1's effect on your response Y will differ depending on the value of another predictor X_2
- Example: back to age of acquisition and word frequency: let's make a median split on AoA and examine frequency effects on lexical decision time:



Interactions between predictors

- However, we want to generalize this so that we don't rely on the median split, but rather on a continuous dependence between the effects of frequency and AoA

Interactions between predictors

- However, we want to generalize this so that we don't rely on the median split, but rather on a continuous dependence between the effects of frequency and AoA
- This is known as an *interaction*

Interactions between predictors

- However, we want to generalize this so that we don't rely on the median split, but rather on a continuous dependence between the effects of frequency and AoA
- This is known as an *interaction*
- Formally, introducing an interaction between predictors X_i and X_j simply means adding a predictor whose value is the *product* of X_i and X_j

$$RT \sim \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \overbrace{X_3}^{=X_1 X_2}$$

Interactions between predictors

- However, we want to generalize this so that we don't rely on the median split, but rather on a continuous dependence between the effects of frequency and AoA
- This is known as an *interaction*
- Formally, introducing an interaction between predictors X_i and X_j simply means adding a predictor whose value is the *product* of X_i and X_j

$$RT \sim \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \overbrace{X_3}^{=X_1 X_2}$$

- Effect on the model matrix for our problem:

$$\mathrm{X} = \begin{bmatrix} 1 & \mathrm{Freq}_1 & \mathrm{AoA}_1 & \mathrm{Freq}_1\mathrm{AoA}_1 \\ 1 & \mathrm{Freq}_2 & \mathrm{AoA}_2 & \mathrm{Freq}_2\mathrm{AoA}_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{Freq}_n & \mathrm{AoA}_n & \mathrm{Freq}_n\mathrm{AoA}_n \end{bmatrix}$$

Interpreting interactions

Resulting model fit:

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\alpha}$	488.55	47.92	$\ll 0.001$
$\widehat{\beta}_{\text{Freq}}$	6.07	5.73	0.29
$\widehat{\beta}_{\text{AoA}}$	33.47	7.35	$\ll 0.001$
$\widehat{\beta}_{\text{Freq}\times\text{AoA}}$	-2.84	0.912	0.00193

- The t statistic for $\widehat{\beta}_{\text{Freq}\times\text{AoA}}$ indicates that the interaction is significant

Interpreting interactions

Resulting model fit:

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\lvert t\rvert)$
$\widehat{\alpha}$	488.55	47.92	$\ll 0.001$
$\widehat{\beta}_{\mathsf{Freq}}$	6.07	5.73	0.29
$\widehat{\beta}_{\mathsf{AoA}}$	33.47	7.35	$\ll 0.001$
$\widehat{\beta}_{\mathsf{Freq}\times\mathsf{AoA}}$	-2.84	0.912	0.00193

- The t statistic for $\widehat{\beta}_{\mathsf{Freq}\times\mathsf{AoA}}$ indicates that the interaction is significant
- Interpretation of the interaction parameter: as AoA increases, the effect of frequency becomes "less positive"

Interpreting interactions

Resulting model fit:

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\alpha}$	488.55	47.92	$\ll 0.001$
$\widehat{\beta}_{\text{Freq}}$	6.07	5.73	0.29
$\widehat{\beta}_{\text{AoA}}$	33.47	7.35	$\ll 0.001$
$\widehat{\beta}_{\text{Freq}\times\text{AoA}}$	-2.84	0.912	0.00193

- The t statistic for $\widehat{\beta}_{\text{Freq}\times\text{AoA}}$ indicates that the interaction is significant
- Interpretation of the interaction parameter: as AoA increases, the effect of frequency becomes “less positive”
- Critically important issue: when there is a higher-order interaction in the model, care must be taken with the interpretation of lower-order terms involved in the interaction

Interpreting interactions

Resulting model fit:

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\alpha}$	488.55	47.92	$\ll 0.001$
$\widehat{\beta}_{\text{Freq}}$	6.07	5.73	0.29
$\widehat{\beta}_{\text{AoA}}$	33.47	7.35	$\ll 0.001$
$\widehat{\beta}_{\text{Freq}\times\text{AoA}}$	-2.84	0.912	0.00193

- The t statistic for $\widehat{\beta}_{\text{Freq}\times\text{AoA}}$ indicates that the interaction is significant
- Interpretation of the interaction parameter: as AoA increases, the effect of frequency becomes “less positive”
- Critically important issue: when there is a higher-order interaction in the model, care must be taken with the interpretation of lower-order terms involved in the interaction
- Here, the “main” effect of frequency ($\widehat{\beta}_{\text{Freq}}$) indicates the effect on RT of increasing word frequency for a hypothetical word with age of acquisition at 0 years!

Interpreting interactions

Resulting model fit:

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\lvert t\rvert)$
$\widehat{\alpha}$	488.55	47.92	$\ll 0.001$
$\widehat{\beta}_{\mathrm{Freq}}$	6.07	5.73	0.29
$\widehat{\beta}_{\mathrm{AoA}}$	33.47	7.35	$\ll 0.001$
$\widehat{\beta}_{\mathrm{Freq}\times\mathrm{AoA}}$	-2.84	0.912	0.00193

- The t statistic for $\widehat{\beta}_{\mathrm{Freq}\times\mathrm{AoA}}$ indicates that the interaction is significant
- Interpretation of the interaction parameter: as AoA increases, the effect of frequency becomes "less positive"
- Critically important issue: when there is a higher-order interaction in the model, care must be taken with the interpretation of lower-order terms involved in the interaction
- Here, the "main" effect of frequency ($\widehat{\beta}_{\mathrm{Freq}}$) indicates the effect on RT of increasing word frequency for a hypothetical word with age of acquisition at 0 years!
- But such words can't exist!

Interpreting interactions

- How do we interpret lower-order terms—e.g., “main effects” of X_1, X_2—in the presence of an interaction term $X_{1\times 2}$?

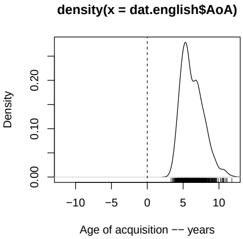
Interpreting interactions

- How do we interpret lower-order terms—e.g., “main effects” of X_1, X_2—in the presence of an interaction term $X_{1\times 2}$?
- One possible answer: we want a “main effect” of X_2 to indicate the effect of X_2 as X_1 remains constant at its average value for the dataset

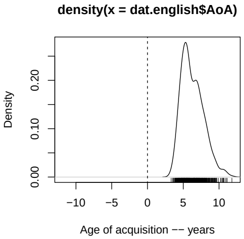
Interpreting interactions

- How do we interpret lower-order terms—e.g., “main effects” of X_1, X_2—in the presence of an interaction term $X_{1\times2}$?
- One possible answer: we want a “main effect” of X_2 to indicate the effect of X_2 as X_1 remains constant at its average value for the dataset
- We can achieve this by centering the predictor(s)

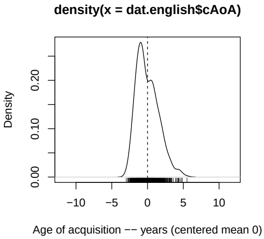
Interpreting interactions—centering predictors



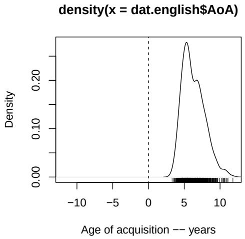
Interpreting interactions—centering predictors



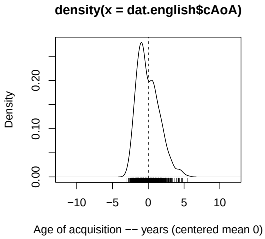
→

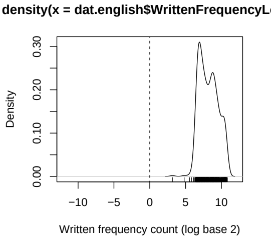


Interpreting interactions—centering predictors

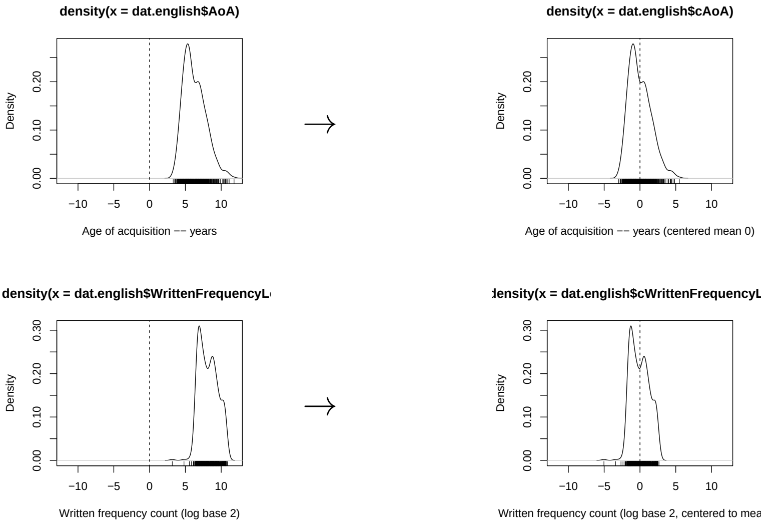


→





Interpreting interactions—centering predictors



Interpreting interactions

With the centered predictors, our model becomes:

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\alpha}$	602.04	1.94	$\ll 0.001$
$\widehat{\beta}_{\text{cFreq}}$	-11.79	1.53	$\ll 0.001$
$\widehat{\beta}_{\text{cAoA}}$	10.14	1.30	$\ll 0.001$
$\widehat{\beta}_{\text{cFreq}\times\text{cAoA}}$	-2.84	0.912	0.00193

Interpreting interactions

With the centered predictors, our model becomes:

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\alpha}$	602.04	1.94	$\ll 0.001$
$\widehat{\beta}_{\text{cFreq}}$	-11.79	1.53	$\ll 0.001$
$\widehat{\beta}_{\text{cAoA}}$	10.14	1.30	$\ll 0.001$
$\widehat{\beta}_{\text{cFreq}\times\text{cAoA}}$	-2.84	0.912	0.00193

- This shows that the "average"-AoA word in our dataset *does* have a substantial frequency effect

Interpreting interactions

With the centered predictors, our model becomes:

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\lvert t\rvert)$
$\widehat{\alpha}$	602.04	1.94	$\ll 0.001$
$\widehat{\beta}_{\text{cFreq}}$	-11.79	1.53	$\ll 0.001$
$\widehat{\beta}_{\text{cAoA}}$	10.14	1.30	$\ll 0.001$
$\widehat{\beta}_{\text{cFreq}\times\text{cAoA}}$	-2.84	0.912	0.00193

- This shows that the "average"-AoA word in our dataset *does* have a substantial frequency effect
- Likewise, the "average"-frequency word in our dataset *does* have an AoA frequency effect

Interpreting interactions

- Even better, however, is *visualizing* the interactions

Interpreting interactions

- Even better, however, is *visualizing* the interactions
- The interaction means that as one predictor changes value, the other predictor's effect changes (and vice versa)

Interpreting interactions

- Even better, however, is *visualizing* the interactions
- The interaction means that as one predictor changes value, the other predictor's effect changes (and vice versa)
- Going back to the regression equation, we can write it as

$$\begin{aligned} RT &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \overbrace{X_3}^{=X_1X_2} + \epsilon \\ &= \alpha + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \epsilon \end{aligned}$$

So

$$\hat{y} = \widehat{\alpha} + \widehat{\beta}_1 X_1 + (\widehat{\beta}_2 + \widehat{\beta}_3 X_1) X_2 + \epsilon$$

Interpreting interactions

- Even better, however, is *visualizing* the interactions
- The interaction means that as one predictor changes value, the other predictor's effect changes (and vice versa)
- Going back to the regression equation, we can write it as

$$\begin{aligned} RT &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \overbrace{X_3}^{=X_1 X_2} + \epsilon \\ &= \alpha + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \epsilon \end{aligned}$$

So

$$\hat{y} = \widehat{\alpha} + \widehat{\beta}_1 X_1 + (\widehat{\beta}_2 + \widehat{\beta}_3 X_1) X_2 + \epsilon$$

- So the estimated effect size of X_2 is $(\widehat{\beta}_2 + \widehat{\beta}_3 X_1)$

Interpreting interactions

- Even better, however, is *visualizing* the interactions
- The interaction means that as one predictor changes value, the other predictor's effect changes (and vice versa)
- Going back to the regression equation, we can write it as

$$\begin{aligned} RT &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \overbrace{X_3}^{=X_1 X_2} + \epsilon \\ &= \alpha + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \epsilon \end{aligned}$$

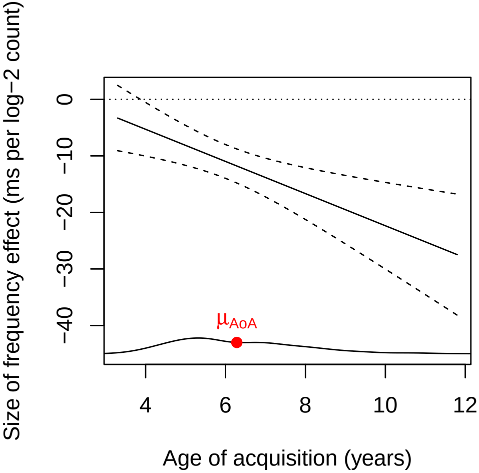
So

$$\hat{y} = \widehat{\alpha} + \widehat{\beta}_1 X_1 + (\widehat{\beta}_2 + \widehat{\beta}_3 X_1) X_2 + \epsilon$$

- So the estimated effect size of X_2 is $(\widehat{\beta}_2 + \widehat{\beta}_3 X_1)$
- Because the estimates $\{\widehat{\beta}_i\}$ are multivariate-normal distributed, we can also use this to compute a standard error (and thus confidence intervals) of the effect size of X_2 for a given value of X_1

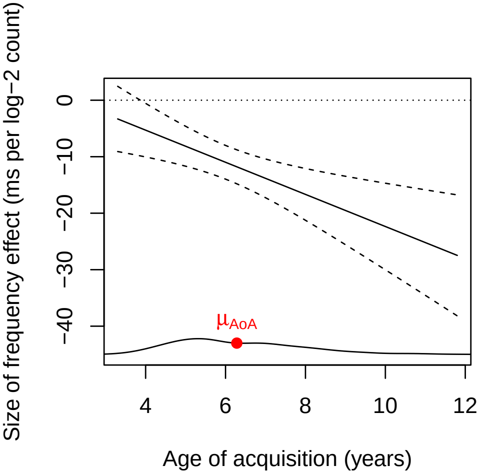
Visualizing interactions

- Let us explore the effect of a word's frequency as a function of its AoA in this model



Visualizing interactions

- Let us explore the effect of a word’s frequency as a function of its AoA in this model



- What are some possible theoretical interpretations of this

Interactions with categorical variables

- Case of an interaction between a k-level categorial variable X_1 and a continuous variable X_2

Interactions with categorical variables

- Case of an interaction between a k-level categorial variable X_1 and a continuous variable X_2
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?

Interactions with categorical variables

- Case of an interaction between a k-level categorial variable X_1 and a continuous variable X_2
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept

Interactions with categorical variables

- Case of an interaction between a k-level categorial variable X_1 and a continuous variable X_2
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept
  - $k - 1$ more for X_1

Interactions with categorical variables

- Case of an interaction between a k-level categorial variable X_1 and a continuous variable X_2
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept
  - $k - 1$ more for X_1
  - 1 more for X_2

Interactions with categorical variables

- Case of an interaction between a k-level categorial variable X_1 and a continuous variable X_2
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept
  - $k - 1$ more for X_1
  - 1 more for X_2
- So, a total of $k + 1$

Interactions with categorical variables

- Case of an interaction between a k-level categorial variable X_1 and a continuous variable X_2
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept
  - $k - 1$ more for X_1
  - 1 more for X_2
- So, a total of $k + 1$
- Adding the interaction between X_1 and X_2 allows for a separate X_2 slope for each value of X_1 (instead of the same slope for all values of X_1)

Interactions with categorical variables

- Case of an interaction between a k-level categorial variable X_1 and a continuous variable X_2
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept
  - $k - 1$ more for X_1
  - 1 more for X_2
- So, a total of $k + 1$
- Adding the interaction between X_1 and X_2 allows for a separate X_2 slope for each value of X_1 (instead of the same slope for all values of X_1)
- Hence this adds a total a total of $k - 1$ new parameters beyond the additive ("main effects") model

Interactions with categorical variables

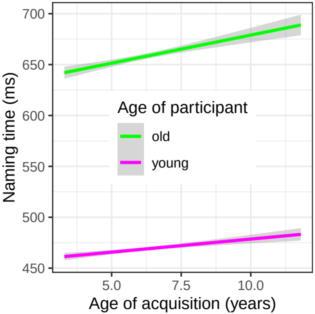
Example: different effects of AoA on naming time for old versus young native English speakers:

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\lvert t\rvert)$
$\widehat{\beta}_{\text{Age=old}}$	623.81	4.81	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young}}$	452.88	4.81	$\ll 0.001$
$\widehat{\beta}_{\text{AoA:Age=old}}$	5.52	0.744	$\ll 0.001$
$\widehat{\beta}_{\text{AoA:Age=young}}$	2.57	0.744	0.000581

Interactions with categorical variables

Example: different effects of AoA on naming time for old versus young native English speakers:

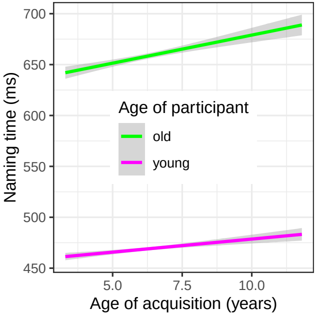
Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\beta}_{\text{Age=old}}$	623.81	4.81	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young}}$	452.88	4.81	$\ll 0.001$
$\widehat{\beta}_{\text{AoA:Age=old}}$	5.52	0.744	$\ll 0.001$
$\widehat{\beta}_{\text{AoA:Age=young}}$	2.57	0.744	0.000581



Interactions with categorical variables

Example: different effects of AoA on naming time for old versus young native English speakers:

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\beta}_{\text{Age=old}}$	623.81	4.81	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young}}$	452.88	4.81	$\ll 0.001$
$\widehat{\beta}_{\text{AoA:Age=old}}$	5.52	0.744	$\ll 0.001$
$\widehat{\beta}_{\text{AoA:Age=young}}$	2.57	0.744	0.000581

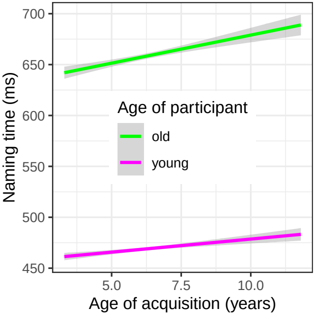


- With this parameterization, the t statistics aren't useful in assessing the strength of evidence for the interaction

Interactions with categorical variables

Example: different effects of AoA on naming time for old versus young native English speakers:

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\beta}_{\text{Age=old}}$	623.81	4.81	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young}}$	452.88	4.81	$\ll 0.001$
$\widehat{\beta}_{\text{AoA:Age=old}}$	5.52	0.744	$\ll 0.001$
$\widehat{\beta}_{\text{AoA:Age=young}}$	2.57	0.744	0.000581



- With this parameterization, the t statistics aren't useful in assessing the strength of evidence for the interaction
- We can use the F-test to do this instead—comparing with a 3-parameter model (sans interaction) gives us $F_{1,1074} = 7.84$, for a p-value of 0.00519

Interactions between categorical variables

- Case of an interaction between a k_1-level categorial variable X_1 and another k_2-level categorial variable X_2, then you have a total of $k_1 k_2$ parameters

Interactions between categorical variables

- Case of an interaction between a k_1-level categorial variable X_1 and another k_2-level categorial variable X_2, then you have a total of $k_1 k_2$ parameters
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?

Interactions between categorical variables

- Case of an interaction between a k_1-level categorial variable X_1 and another k_2-level categorial variable X_2, then you have a total of $k_1 k_2$ parameters
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept

Interactions between categorical variables

- Case of an interaction between a k_1-level categorial variable X_1 and another k_2-level categorial variable X_2, then you have a total of $k_1 k_2$ parameters
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept
  - $k_1 - 1$ more for X_1

Interactions between categorical variables

- Case of an interaction between a k_1-level categorial variable X_1 and another k_2-level categorial variable X_2, then you have a total of $k_1 k_2$ parameters
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept
  - $k_1 - 1$ more for X_1
  - $k_2 - 1$ more for X_2

Interactions between categorical variables

- Case of an interaction between a k_1-level categorial variable X_1 and another k_2-level categorial variable X_2, then you have a total of $k_1 k_2$ parameters
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept
  - $k_1 - 1$ more for X_1
  - $k_2 - 1$ more for X_2
- So, a total of $k_1 + k_2 - 1$

Interactions between categorical variables

- Case of an interaction between a k_1-level categorial variable X_1 and another k_2-level categorial variable X_2, then you have a total of $k_1 k_2$ parameters
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept
  - $k_1 - 1$ more for X_1
  - $k_2 - 1$ more for X_2
- So, a total of $k_1 + k_2 - 1$
- Adding the interaction between X_1 and X_2 means that there is a separate effect of each logically possible combination of the two predictors

Interactions between categorical variables

- Case of an interaction between a k_1-level categorial variable X_1 and another k_2-level categorial variable X_2, then you have a total of $k_1 k_2$ parameters
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept
  - $k_1 - 1$ more for X_1
  - $k_2 - 1$ more for X_2
- So, a total of $k_1 + k_2 - 1$
- Adding the interaction between X_1 and X_2 means that there is a separate effect of each logically possible combination of the two predictors
- So there is a total of $k_1 k_2$ parameters (one for the intercept and $k_1 k_2 - 1$ for the overall interaction)

Interactions between categorical variables

- Case of an interaction between a k_1-level categorial variable X_1 and another k_2-level categorial variable X_2, then you have a total of $k_1 k_2$ parameters
- Review: how many parameters does an additive model of $Y \sim X_1 + X_2$ have?
  - 1 for the intercept
  - $k_1 - 1$ more for X_1
  - $k_2 - 1$ more for X_2
- So, a total of $k_1 + k_2 - 1$
- Adding the interaction between X_1 and X_2 means that there is a separate effect of each logically possible combination of the two predictors
- So there is a total of $k_1 k_2$ parameters (one for the intercept and $k_1 k_2 - 1$ for the overall interaction)
- Hence this adds a total a total of $k_1 k_2 - k_1 - k_2 + 1$ new parameters beyond the additive ("main effects") model

Interactions between categorical variables

Example: the relationship between initial phone type and native speaker age on naming times

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\beta}_{\text{Age=old,Init=burst}}$	654.83	1.66	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=burst}}$	469.65	1.66	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=frication}}$	669.72	1.84	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=frication}}$	475.77	1.84	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=other}}$	643.18	2.82	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=other}}$	450.76	2.82	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=vowel}}$	653.65	8.21	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=vowel}}$	474.00	8.21	$\ll 0.001$

- With this parameterization the model is just encoding eight different means

Interactions between categorical variables

Example: the relationship between initial phone type and native speaker age on naming times

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\lvert t\rvert)$
$\widehat{\beta}_{\text{Age=old,Init=burst}}$	654.83	1.66	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=burst}}$	469.65	1.66	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=frication}}$	669.72	1.84	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=frication}}$	475.77	1.84	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=other}}$	643.18	2.82	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=other}}$	450.76	2.82	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=vowel}}$	653.65	8.21	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=vowel}}$	474.00	8.21	$\ll 0.001$

- With this parameterization the model is just encoding eight different means
- Once again, we can use the F test to compare this model with an additive model of parameters, giving us

Interactions between categorical variables

Example: the relationship between initial phone type and native speaker age on naming times

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\beta}_{\text{Age=old,Init=burst}}$	654.83	1.66	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=burst}}$	469.65	1.66	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=frication}}$	669.72	1.84	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=frication}}$	475.77	1.84	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=other}}$	643.18	2.82	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=other}}$	450.76	2.82	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=vowel}}$	653.65	8.21	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=vowel}}$	474.00	8.21	$\ll 0.001$

- With this parameterization the model is just encoding eight different means
- Once again, we can use the F test to compare this model with an additive model of 5 parameters, giving us

Interactions between categorical variables

Example: the relationship between initial phone type and native speaker age on naming times

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\beta}_{\text{Age=old,Init=burst}}$	654.83	1.66	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=burst}}$	469.65	1.66	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=frication}}$	669.72	1.84	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=frication}}$	475.77	1.84	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=other}}$	643.18	2.82	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=other}}$	450.76	2.82	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=vowel}}$	653.65	8.21	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=vowel}}$	474.00	8.21	$\ll 0.001$

- With this parameterization the model is just encoding eight different means
- Once again, we can use the F test to compare this model with an additive model of 5 parameters, giving us

Interactions between categorical variables

Example: the relationship between initial phone type and native speaker age on naming times

Predictor	$\widehat{\beta}$	$SE(\widehat{\beta})$	$p(>\|t\|)$
$\widehat{\beta}_{\text{Age=old,Init=burst}}$	654.83	1.66	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=burst}}$	469.65	1.66	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=frication}}$	669.72	1.84	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=frication}}$	475.77	1.84	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=other}}$	643.18	2.82	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=other}}$	450.76	2.82	$\ll 0.001$
$\widehat{\beta}_{\text{Age=old,Init=vowel}}$	653.65	8.21	$\ll 0.001$
$\widehat{\beta}_{\text{Age=young,Init=vowel}}$	474.00	8.21	$\ll 0.001$

- With this parameterization the model is just encoding eight different means
- Once again, we can use the F test to compare this model with an additive model of 5 parameters, giving us $F_{3,1070} = 2.52$, for a p-value of 0.0563

General words of wisdom for categorical factors & interactions

- There are many different ways to parameterize any model

General words of wisdom for categorical factors & interactions

- There are many different ways to parameterize any model
- It's easy to get lost in the apparent cleverness of setting model contrasts for interesting parameterizations

General words of wisdom for categorical factors & interactions

- There are many different ways to parameterize any model
- It's easy to get lost in the apparent cleverness of setting model contrasts for interesting parameterizations
- Don't lose sight of the more important issue: model class

General words of wisdom for categorical factors & interactions

- There are many different ways to parameterize any model
- It's easy to get lost in the apparent cleverness of setting model contrasts for interesting parameterizations
- Don't lose sight of the more important issue: model class
- Because we're operating under maximum likelihood, if one model is a reparameterization of another, they make the same predictions and yield the same inferences for model comparison!

General words of wisdom for categorical factors & interactions

- There are many different ways to parameterize any model
- It's easy to get lost in the apparent cleverness of setting model contrasts for interesting parameterizations
- Don't lose sight of the more important issue: model class
- Because we're operating under maximum likelihood, if one model is a reparameterization of another, they make the same predictions and yield the same inferences for model comparison!
- Model comparisons (for simple linear regression, the F test; more generally, the likelihood ratio test) are your friend!

Dichotomous categorical response variables

- We have generalized linear regression to categorical *predictor* variables

Dichotomous categorical response variables

- We have generalized linear regression to categorical *predictor* variables
- However, we have not yet addressed the case when the *response* variable is categorical

Dichotomous categorical response variables

- We have generalized linear regression to categorical *predictor* variables
- However, we have not yet addressed the case when the *response* variable is categorical
- Let's consider the case of a dichotomous response variable

Dichotomous categorical response variables

- We have generalized linear regression to categorical *predictor* variables
- However, we have not yet addressed the case when the *response* variable is categorical
- Let's consider the case of a dichotomous response variable
- Example: the dative alternation (Bresnan et al., 2007)

Sally sent [the children]$_{\text{Recip}}$ [toys]$_{\text{Theme}}$	**Double Object**
Sally sent [toys]$_{\text{Theme}}$ to [the children]$_{\text{Recip}}$	**Prepositional Object**

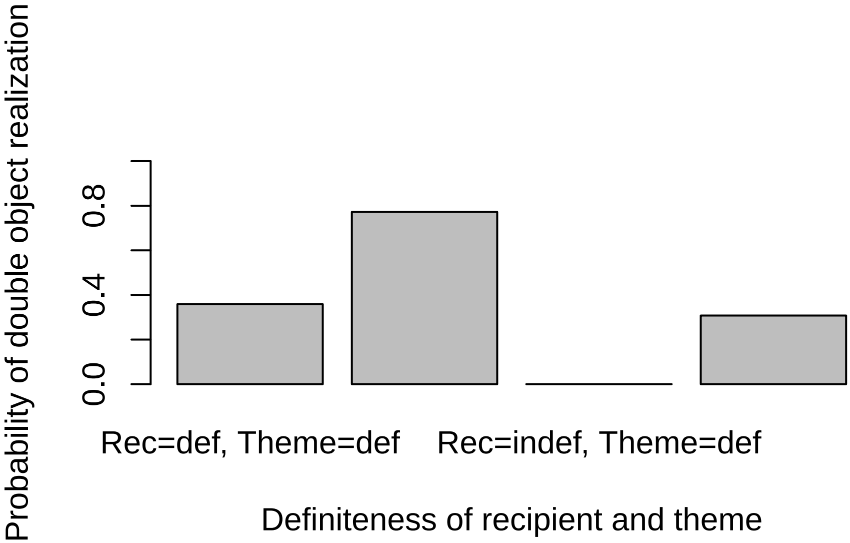
Dichotomous categorical response variables

- We have generalized linear regression to categorical *predictor* variables
- However, we have not yet addressed the case when the *response* variable is categorical
- Let's consider the case of a dichotomous response variable
- Example: the dative alternation (Bresnan et al., 2007)

Sally sent [the children]$_{\text{Recip}}$ [toys]$_{\text{Theme}}$	**Double Object**
Sally sent [toys]$_{\text{Theme}}$ to [the children]$_{\text{Recip}}$	**Prepositional Object**

- We looked briefly before at the effects of definiteness of the theme (*toys/the toys*) and recipient (*children/the children*)

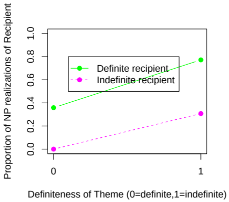
Dichotomous categorical responses



- We could learn these four separate means, but we would fail to capture the systematicity of the effects seen here

Dichotomous categorical responses

Another way of representing the four means:



- This looks like what we called an *additive pattern* for linear regression

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where X_1 is 1 iff the theme is indefinite, and X_2 is 1 iff the recipient is indefinite (both 0 otherwise)

Problems for linear models with categorical response

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ $X_1 = 1$ iff theme indefinite, $X_2 = 1$ iff recipient indefinite (both 0 otherwise)

1. Bad predictions for individual observations: in linear regression, the noise term ϵ is *Gaussian* (normally distributed)—it predicts that any continuous value is possible

Problems for linear models with categorical response

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ $X_1 = 1$ iff theme indefinite, $X_2 = 1$ iff recipient indefinite (both 0 otherwise)

1. Bad predictions for individual observations: in linear regression, the noise term ϵ is *Gaussian* (normally distributed)—it predicts that any continuous value is possible
   - The only really possible outcomes for individual observations are 0 (PP recipient) and 1 (NP recipient)

Problems for linear models with categorical response

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ $X_1 = 1$ iff theme indefinite, $X_2 = 1$ iff recipient indefinite (both 0 otherwise)

1. Bad predictions for individual observations: in linear regression, the noise term ϵ is *Gaussian* (normally distributed)—it predicts that any continuous value is possible
   - The only really possible outcomes for individual observations are 0 (PP recipient) and 1 (NP recipient)
   - Remember that our observed "means" are averages of many 0 and 1 observations!

```
##                            Definiteness of recipient and theme
## Realization     Rec=def, Theme=def Rec=def, Theme=indef
##     Double obj.                 19                   78
##     Prep. obj.                  34                   23
##                            Definiteness of recipient and theme
## Realization     Rec=indef, Theme=def Rec=indef, Theme=indef
##     Double obj.                    0                      4
##     Prep. obj.                     5                      9
```

Problems for linear models with categorical response

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ $X_1 = 1$ iff theme indefinite, $X_2 = 1$ iff recipient indefinite (both 0 otherwise)

1. Bad predictions for individual observations: in linear regression, the noise term ϵ is *Gaussian* (normally distributed)—it predicts that any continuous value is possible
   - The only really possible outcomes for individual observations are 0 (PP recipient) and 1 (NP recipient)
   - Remember that our observed "means" are averages of many 0 and 1 observations!

```
##                             Definiteness of recipient and theme
## Realization      Rec=def, Theme=def Rec=def, Theme=indef
##     Double obj.                  19                   78
##     Prep. obj.                   34                   23
##                             Definiteness of recipient and theme
## Realization      Rec=indef, Theme=def Rec=indef, Theme=indef
##     Double obj.                    0                     4
##     Prep. obj.                     5                     9
```

2. Bad predicted means: in linear regression, there is no guarantee that the predicted mean response $\hat{y}$ will fall between 0 and 1, even if all individual observations fall within this range

Bad predicted means with linear regression for categorical response

- Consider a case where a predictor is continuous and the response is categorical:

Recipient is NP	Recipient is PP
Mary sent John a shiny toy	Mary sent a shiny toy to John
Mary sent her friend a shiny toy	Mary sent a shiny toy to her friend
Mary sent every kid in the room a shiny toy	Mary sent a shiny toy to every kid in the room

Bad predicted means with linear regression for categorical response

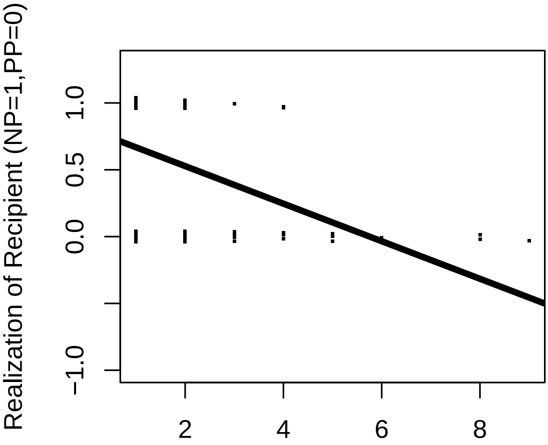
- Consider a case where a predictor is continuous and the response is categorical:

Recipient is NP	Recipient is PP
Mary sent John a shiny toy	Mary sent a shiny toy to John
Mary sent her friend a shiny toy	Mary sent a shiny toy to her friend
Mary sent every kid in the room a shiny toy	Mary sent a shiny toy to every kid in the room

- We could quantify the *size* of the recipient in any number of ways (here we'll use length in # of words)

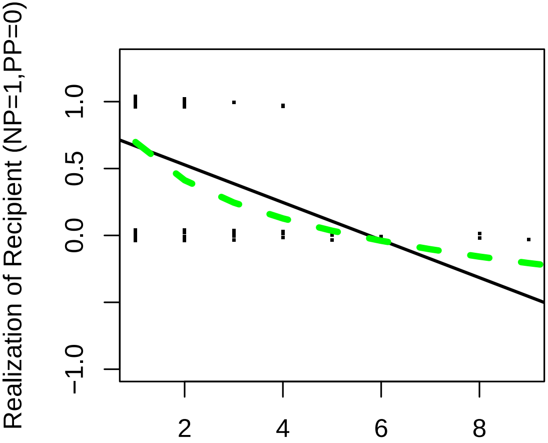
Dichotomous categorical response variables

Here's what happens when we learn a linear regression model on recipient length:



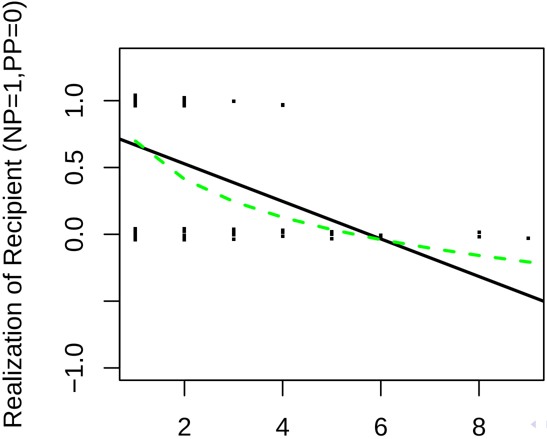
Bad predicted means with linear regression for categorical response

Same problem if we use *log* of recipient length as a predictor:



Bad predicted means with linear regression for categorical response

Even spline-based methods (which we haven't covered yet) give us the same problem, too, at the far end of the range of lengths:



Problems with linear regression for categorical response

- So linear regression is bad for categorical response variables in:

Problems with linear regression for categorical response

- So linear regression is bad for categorical response variables in:
    1. The noise distribution it assumes around the predicted mean

Problems with linear regression for categorical response

- So linear regression is bad for categorical response variables in:
  1. The noise distribution it assumes around the predicted mean
  2. The range of the predicted mean allowed

Problems with linear regression for categorical response

- So linear regression is bad for categorical response variables in:
  1. The noise distribution it assumes around the predicted mean
  2. The range of the predicted mean allowed
- Fortunately, the framework of generalized linear models (GLMs) gives us the flexibility to deal with these problems!

Reviewing GLMs

Assumptions of the generalized linear model (GLM):

1. Predictors $\{X_i\}$ influence Y through the mediation of a linear predictor η;

Reviewing GLMs

Assumptions of the generalized linear model (GLM):

1. Predictors $\{X_i\}$ influence Y through the mediation of a linear predictor η;
2. η is a linear combination of the $\{X_i\}$:

Reviewing GLMs

Assumptions of the generalized linear model (GLM):

1. Predictors $\{X_i\}$ influence Y through the mediation of a linear predictor η;
2. η is a linear combination of the $\{X_i\}$:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \qquad \text{(linear predictor)}$$

Reviewing GLMs

Assumptions of the generalized linear model (GLM):

1. Predictors $\{X_i\}$ influence Y through the mediation of a linear predictor η;
2. η is a linear combination of the $\{X_i\}$:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \qquad \text{(linear predictor)}$$

3. η determines the predicted mean μ of Y

$$\eta = l(\mu) \qquad \text{(link function)}$$

Reviewing GLMs

Assumptions of the generalized linear model (GLM):

1. Predictors $\{X_i\}$ influence Y through the mediation of a linear predictor η;
2. η is a linear combination of the $\{X_i\}$:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \qquad \text{(linear predictor)}$$

3. η determines the predicted mean μ of Y

$$\eta = l(\mu) \qquad \text{(link function)}$$

4. There is some noise distribution of Y around the predicted mean μ of Y:

$$P(Y = y; \mu)$$

Logit GLMs for dichotomous responses

- Choosing a different link function and noise distribution gives us the logit model

Logit GLMs for dichotomous responses

- Choosing a different link function and noise distribution gives us the logit model
- Logit link function:

$$\eta = \log \frac{\mu}{1 - \mu}$$

Logit GLMs for dichotomous responses

- Choosing a different link function and noise distribution gives us the logit model
- Logit link function:

$$\eta = \log \frac{\mu}{1-\mu}$$

- Bernoulli noise distribution around predicted mean μ:

$$P(Y=y|\mu) = \begin{cases} \mu & y = 1 \\ 1-\mu & y = 0 \\ 0 & \text{otherwise} \end{cases}$$

Logit GLMs for dichotomous responses

- Choosing a different link function and noise distribution gives us the logit model
- Logit link function:

$$\eta = \log \frac{\mu}{1-\mu}$$

- Bernoulli noise distribution around predicted mean μ:

$$P(Y = y|\mu) = \begin{cases} \mu & y = 1 \\ 1-\mu & y = 0 \\ 0 & \text{otherwise} \end{cases}$$

- The linear predictor remains as it was before:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m$$

Logit GLMs for dichotomous responses

- Choosing a different link function and noise distribution gives us the logit model
- Logit link function:

$$\eta = \log \frac{\mu}{1-\mu}$$

- Bernoulli noise distribution around predicted mean μ:

$$P(Y = y|\mu) = \begin{cases} \mu & y = 1 \\ 1-\mu & y = 0 \\ 0 & \text{otherwise} \end{cases}$$

- The linear predictor remains as it was before:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m$$

- Using logit GLMs to fit data with dichotomous response variables is called logistic regression

The logit link function

- A lot of the action in logistic regression is in the logit link function:

$$\eta = \log \frac{\mu}{1-\mu}$$

The logit link function

- A lot of the action in logistic regression is in the logit link function:

$$\eta = \log \frac{\mu}{1-\mu}$$

- Looking at its *inverse* is equally useful:

$$\mu = \frac{e^{\eta}}{1+e^{\eta}}$$

The logit link function

- A lot of the action in logistic regression is in the logit link function:

$$\eta = \log \frac{\mu}{1-\mu}$$

- Looking at its *inverse* is equally useful:

$$\mu = \frac{e^{\eta}}{1+e^{\eta}}$$

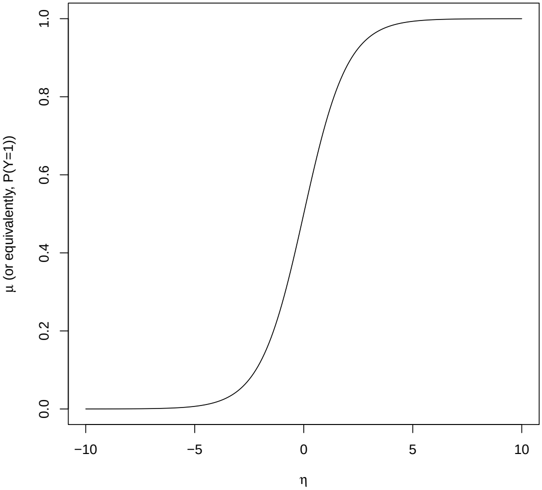
The logit link function

- A lot of the action in logistic regression is in the logit link function:

$$\eta = \log \frac{\mu}{1-\mu}$$

- Looking at its *inverse* is equally useful:

$$\mu = \frac{e^\eta}{1+e^\eta}$$



Estimating parameters in logistic regression

$$\begin{aligned} \eta &= \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \\ \mu &= \frac{e^\eta}{1 + e^\eta} \\ P(Y = y|\mu) &= \begin{cases} \mu & y = 1 \\ 1 - \mu & y = 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

- As with linear regression, the regression weights $\langle \alpha, \beta_1, \dots, \beta_m \rangle$ must be learned

Estimating parameters in logistic regression

$$\boxed{\begin{aligned}\eta &= \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \\ \mu &= \frac{e^\eta}{1+e^\eta} \\ P(Y=y|\mu) &= \begin{cases} \mu & y=1 \\ 1-\mu & y=0 \\ 0 & \text{otherwise} \end{cases}\end{aligned}}$$

- As with linear regression, the regression weights $\langle \alpha, \beta_1, \ldots, \beta_m \rangle$ must be learned
- Unlike linear regression, there is no additional noise parameter to be learned (σ^2 in linear regression)

Estimating parameters in logistic regression

$$\begin{aligned}
\eta &= \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \\
\mu &= \frac{e^\eta}{1 + e^\eta} \\
P(Y = y|\mu) &= \begin{cases} \mu & y = 1 \\ 1 - \mu & y = 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

- As with linear regression, the regression weights $\langle \alpha, \beta_1, \ldots, \beta_m \rangle$ must be learned
- Unlike linear regression, there is no additional noise parameter to be learned (σ^2 in linear regression)
- Once again, we can use the method of maximum likelihood (which we use in the example here) or Bayesian inference to estimate parameters

Interpreting an additive logistic regression model

- Here's a logistic regression model for additive effects of theme and recipient definiteness:

$$\eta = \alpha + \beta_{\text{ThemeDef}} X_{\text{ThemeDef}} + \beta_{\text{RecDef}} X_{\text{RecDef}}$$

$$\mu = \frac{e^{\eta}}{1 + e^{\eta}}$$

$$P(Y = y|\mu) = \begin{cases} \mu & y = 1 \\ 1 - \mu & y = 0 \\ 0 & \text{otherwise} \end{cases}$$

Interpreting an additive logistic regression model

- Here's a logistic regression model for additive effects of theme and recipient definiteness:

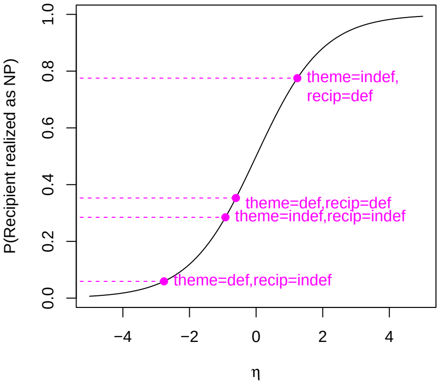
$$
\begin{aligned}
\eta &= \alpha + \beta_{\text{ThemeDef}} X_{\text{ThemeDef}} + \beta_{\text{RecDef}} X_{\text{RecDef}} \\
\mu &= \frac{e^{\eta}}{1 + e^{\eta}} \\
P(Y = y|\mu) &= \begin{cases} \mu & y = 1 \\ 1 - \mu & y = 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

- The maximum likelihood estimate for the three regression parameters is

$\widehat{\alpha}$	-0.61
$\widehat{\beta}_{\text{ThemeDef}}$	1.8
$\widehat{\beta}_{\text{RecDef}}$	-2.2

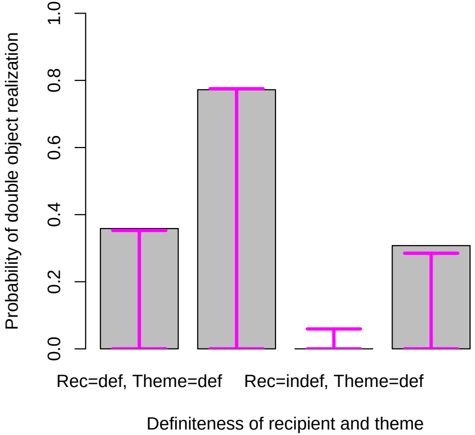
Interpreting an additive logistic regression model

$\widehat{\alpha}$	-0.61
$\widehat{\beta}_{\text{ThemeDef}}$	1.8
$\widehat{\beta}_{\text{RecDef}}$	-2.2



Interpreting an additive logistic regression model

This additive model does a decent job of modeling the true means!



Confidence regions for logistic regression

- Let us write the linear-predictor part of our GLM in matrix form:

$$\boldsymbol{\eta} = \mathsf{X}\boldsymbol{\beta}$$

Confidence regions for logistic regression

- Let us write the linear-predictor part of our GLM in matrix form:

$$\boldsymbol{\eta} = \mathsf{X}\boldsymbol{\beta}$$

- In linear regression, we built confidence regions for parameter estimates on the basis that the covariance matrix of the MLE $\widehat{\boldsymbol{\beta}}$ can be written *exactly* as

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathsf{X}^\mathsf{T}\mathsf{X})^{-1}$$

Confidence regions for logistic regression

- Let us write the linear-predictor part of our GLM in matrix form:

$$\boldsymbol{\eta} = \mathsf{X}\boldsymbol{\beta}$$

- In linear regression, we built confidence regions for parameter estimates on the basis that the covariance matrix of the MLE $\widehat{\boldsymbol{\beta}}$ can be written *exactly* as

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathsf{X}^{\mathsf{T}}\mathsf{X})^{-1}$$

- In logistic regression and other GLMs, we make confidence regions and model comparisons based on constructs whose *asymptotic* (=approximately true, and increasingly accurate as sample sizes increase) form we can state

Confidence regions for logistic regression

- Let us write the linear-predictor part of our GLM in matrix form:

$$\boldsymbol{\eta} = \mathsf{X}\boldsymbol{\beta}$$

- In linear regression, we built confidence regions for parameter estimates on the basis that the covariance matrix of the MLE $\widehat{\boldsymbol{\beta}}$ can be written *exactly* as

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathsf{X}^{\mathsf{T}}\mathsf{X})^{-1}$$

- In logistic regression and other GLMs, we make confidence regions and model comparisons based on constructs whose *asymptotic* (=approximately true, and increasingly accurate as sample sizes increase) form we can state
- For confidence regions: asymptotically, the covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} \frac{\partial^2 L(\beta_1)}{\partial \beta_1^2} & \frac{\partial^2 L(\beta_2)}{\partial \beta_1 \beta_2} & \cdots & \frac{\partial^2 L(\beta_m)}{\partial \beta_1 \beta_m} \\ \frac{\partial^2 L(\beta_1)}{\partial \beta_1 \beta_2} & \frac{\partial^2 L(\beta_2)}{\partial \beta_2^2} & \cdots & \frac{\partial^2 L(\beta_m)}{\partial \beta_2 \beta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 L(\beta_1)}{\partial \beta_1 \beta_m} & \frac{\partial^2 L(\beta_2)}{\partial \beta_2 \beta_m} & \cdots & \frac{\partial^2 L(\beta_m)}{\partial \beta_m^2} \end{bmatrix}$$

(when certain regularity conditions hold). This is known as the Fisher information matrix.

Confidence regions for logistic regression

- A confidence region for predictors of a model estimated under maximum likelihood can be constructed similarly to the case in linear regression: for any size-k subset of predictors $\boldsymbol{\beta}'$, the quantity

$$(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})^T \left(\mathrm{Cov}(\widehat{\boldsymbol{\beta}}')\right)^{-1} (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})^T$$

(a multivariate Wald statistic) is *asymptotically* χ^2_k distributed.

Confidence regions for logistic regression

- A confidence region for predictors of a model estimated under maximum likelihood can be constructed similarly to the case in linear regression: for any size-k subset of predictors $\boldsymbol{\beta}'$, the quantity

$$(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})^T \left(\mathrm{Cov}(\widehat{\boldsymbol{\beta}}')\right)^{-1} (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})^T$$

  (a multivariate Wald statistic) is *asymptotically* χ^2_k distributed.
- For a single model parameter β, we can equivalently say that

$$\frac{(\widehat{\beta} - \beta)}{SE(\widehat{\beta})}$$

  is asymptotically normally distributed, where $SE(\widehat{\beta}) = \sqrt{\mathrm{Var}(\widehat{\boldsymbol{\beta}})}$. This quantity for $\beta = 0$ is often called the Wald z-statistic.

Confidence regions for logistic regression

- A confidence region for predictors of a model estimated under maximum likelihood can be constructed similarly to the case in linear regression: for any size-k subset of predictors $\boldsymbol{\beta}'$, the quantity

$$(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})^T \left(\mathrm{Cov}(\widehat{\boldsymbol{\beta}}')\right)^{-1} (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})^T$$

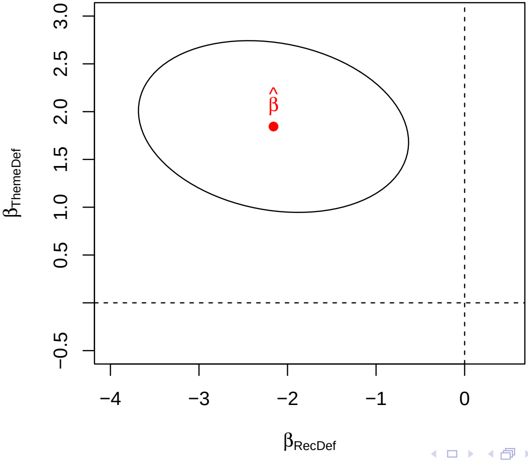
  (a multivariate Wald statistic) is *asymptotically* χ^2_k distributed.
- For a single model parameter β, we can equivalently say that

$$\frac{(\widehat{\beta} - \beta)}{SE(\widehat{\beta})}$$

  is asymptotically normally distributed, where $SE(\widehat{\beta}) = \sqrt{\mathrm{Var}(\widehat{\boldsymbol{\beta}})}$. This quantity for $\beta = 0$ is often called the Wald *z*-statistic.
- Caution! These approximations break down when the estimates $\widehat{\beta}$ are large—most notably, when a single predictor allows *perfect* prediction of an outcome (always 0, or always 1)

Confidence regions in logistic regression

For example, a confidence region for the effects of recipient and theme definiteness:



Interactions in logistic regression

- Interactions work exactly the same for all GLMs, including logistic regression, as for linear regression

Interactions in logistic regression

- Interactions work exactly the same for all GLMs, including logistic regression, as for linear regression
- Critically, the interaction terms go into the equation for the linear predictor

$$\begin{aligned}\eta = \alpha &+ \beta_{\mathsf{RecDef}} X_{\mathsf{RecDef}} \\ &+ \beta_{\mathsf{ThemeDef}} X_{\mathsf{ThemeDef}} \\ &+ \beta_{\mathsf{RecDef:ThemeDef}} X_{\mathsf{RecDef}} X_{\mathsf{ThemeDef}}\end{aligned}$$

Interactions in logistic regression

- Interactions work exactly the same for all GLMs, including logistic regression, as for linear regression
- Critically, the interaction terms go into the equation for the linear predictor

$$\begin{aligned}\eta = \alpha &+ \beta_{\text{RecDef}} X_{\text{RecDef}} \\ &+ \beta_{\text{ThemeDef}} X_{\text{ThemeDef}} \\ &+ \beta_{\text{RecDef:ThemeDef}} X_{\text{RecDef}} X_{\text{ThemeDef}}\end{aligned}$$

- Crucial to remember the coding scheme for these categorical predictors—here we'll stay with $X_{\text{ThemeDef}} = 1$ iff theme indefinite, $X_{\text{RecDef}} = 1$ iff recipient indefinite (both 0 otherwise)

Interactions in logistic regression

- MLE fit of the with-interactions model for the *send* data:

$\widehat{\alpha}$	-0.5819215
$\widehat{\beta}_{\text{RecDef}}$	-16
$\widehat{\beta}_{\text{ThemeDef}}$	1.803136
$\widehat{\beta}_{\text{RecDef:ThemeDef}}$	13.952

- However, the standard error of $\widehat{\beta}_{\text{RecDef:ThemeDef}}$ is huge: 1151563

Interactions in logistic regression

- MLE fit of the with-interactions model for the *send* data:

$\widehat{\alpha}$	-0.5819215
$\widehat{\beta}_{\text{RecDef}}$	-16
$\widehat{\beta}_{\text{ThemeDef}}$	1.803136
$\widehat{\beta}_{\text{RecDef:ThemeDef}}$	13.952

- However, the standard error of $\widehat{\beta}_{\text{RecDef:ThemeDef}}$ is huge: 1151563
- This ultimately arose because there was a *perfect prediction* possible:

```
##                            Definiteness of recipient and theme
## Realization    Rec=indef, Theme=def
##    Double obj.                    0
##    Prep. obj.                     5
```

Interactions in logistic regression

- MLE fit of the with-interactions model for the *send* data:

$\widehat{\alpha}$	-0.5819215
$\widehat{\beta}_{\text{RecDef}}$	-16
$\widehat{\beta}_{\text{ThemeDef}}$	1.803136
$\widehat{\beta}_{\text{RecDef:ThemeDef}}$	13.952

- However, the standard error of $\widehat{\beta}_{\text{RecDef:ThemeDef}}$ is huge: 1151563
- This ultimately arose because there was a *perfect prediction* possible:

```
##                     Definiteness of recipient and theme
## Realization    Rec=indef, Theme=def
##    Double obj.                    0
##    Prep. obj.                     5
```

- Remember, in these situations you cannot trust the Wald z-statistic $\left(\frac{\widehat{\beta}}{SE(\widehat{\beta})}\right)$!

The likelihood ratio test

- For linear regression, hypothesis testing for a single model parameter using the t-statistic yielded *exactly* the same result as explicit model comparison with the F-statistic

The likelihood ratio test

- For linear regression, hypothesis testing for a single model parameter using the t-statistic yielded *exactly* the same result as explicit model comparison with the F-statistic
- This is a special property of *linear* regression, and does *not* generalize to other GLMs (or to the more complex models we'll see further down the line)

The likelihood ratio test

- For linear regression, hypothesis testing for a single model parameter using the t-statistic yielded *exactly* the same result as explicit model comparison with the F-statistic
- This is a special property of *linear* regression, and does *not* generalize to other GLMs (or to the more complex models we'll see further down the line)
- Instead, the more general method for hypothesis testing is the likelihood ratio test

The likelihood ratio test

- For linear regression, hypothesis testing for a single model parameter using the t-statistic yielded *exactly* the same result as explicit model comparison with the F-statistic
- This is a special property of *linear* regression, and does *not* generalize to other GLMs (or to the more complex models we'll see further down the line)
- Instead, the more general method for hypothesis testing is the likelihood ratio test
- We saw this before, in the end of the chapter on frequentist hypothesis testing

The likelihood ratio test

- For *nested* models $M_0 \subset M_A$ with k_0 and k_A free parameters respectively, the statistic

$$-2 \log \frac{\max \mathrm{Lik}_{M_0}(\mathrm{y})}{\max \mathrm{Lik}_{M_A}(\mathrm{y})}$$

is (asymptotically, in the limit of large data) distributed as $\chi^2_{k_A - k_0}$ if M_0 is true

The likelihood ratio test

- For *nested* models $M_0 \subset M_A$ with k_0 and k_A free parameters respectively, the statistic

$$-2 \log \frac{\max \mathrm{Lik}_{M_0}(\mathrm{y})}{\max \mathrm{Lik}_{M_A}(\mathrm{y})}$$

  is (asymptotically, in the limit of large data) distributed as $\chi^2_{k_A - k_0}$ if M_0 is true
- This statistic doesn't have the same problems that the Wald z statistic has, so it can be used very generally to compare nested models

The likelihood ratio test

- For *nested* models $M_0 \subset M_A$ with k_0 and k_A free parameters respectively, the statistic

$$-2 \log \frac{\max \mathrm{Lik}_{M_0}(\mathrm{y})}{\max \mathrm{Lik}_{M_A}(\mathrm{y})}$$

  is (asymptotically, in the limit of large data) distributed as $\chi^2_{k_A - k_0}$ if M_0 is true
- This statistic doesn't have the same problems that the Wald z statistic has, so it can be used very generally to compare nested models
- In our case, the additive model for recipient and theme animacy had log-likelihood of -97.1, whereas the interactive model had log-likelihood of -96.8

The likelihood ratio test

- For *nested* models $M_0 \subset M_A$ with k_0 and k_A free parameters respectively, the statistic

$$-2 \log \frac{\max \mathrm{Lik}_{M_0}(\mathrm{y})}{\max \mathrm{Lik}_{M_A}(\mathrm{y})}$$

  is (asymptotically, in the limit of large data) distributed as $\chi^2_{k_A - k_0}$ if M_0 is true
- This statistic doesn't have the same problems that the Wald z statistic has, so it can be used very generally to compare nested models
- In our case, the additive model for recipient and theme animacy had log-likelihood of -97.1, whereas the interactive model had log-likelihood of -96.8
- They differed in 1 parameter, and the cumulative distribution function of χ^2_1 for 0.66 is 0.582, so we conclude that the interaction is statistically

The likelihood ratio test

- For *nested* models $M_0 \subset M_A$ with k_0 and k_A free parameters respectively, the statistic

$$-2 \log \frac{\max \mathrm{Lik}_{M_0}(\mathrm{y})}{\max \mathrm{Lik}_{M_A}(\mathrm{y})}$$

  is (asymptotically, in the limit of large data) distributed as $\chi^2_{k_A - k_0}$ if M_0 is true
- This statistic doesn't have the same problems that the Wald z statistic has, so it can be used very generally to compare nested models
- In our case, the additive model for recipient and theme animacy had log-likelihood of -97.1, whereas the interactive model had log-likelihood of -96.8
- They differed in 1 parameter, and the cumulative distribution function of χ^2_1 for 0.66 is 0.582, so we conclude that the interaction is statistically

The likelihood ratio test

- For *nested* models $M_0 \subset M_A$ with k_0 and k_A free parameters respectively, the statistic

$$-2 \log \frac{\max \mathrm{Lik}_{M_0}(\mathrm{y})}{\max \mathrm{Lik}_{M_A}(\mathrm{y})}$$

  is (asymptotically, in the limit of large data) distributed as $\chi^2_{k_A - k_0}$ if M_0 is true
- This statistic doesn't have the same problems that the Wald z statistic has, so it can be used very generally to compare nested models
- In our case, the additive model for recipient and theme animacy had log-likelihood of -97.1, whereas the interactive model had log-likelihood of -96.8
- They differed in 1 parameter, and the cumulative distribution function of χ^2_1 for 0.66 is 0.582, so we conclude that the interaction is statistically insignificant

References I

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, H. (2007). Predicting the dative alternation. In Boume, G., Kraemer, I., and Zwarts, J., editors, *Cognitive Foundations of Interpretation*, pages 69–95. Amsterdam: Royal Netherlands Academy of Science.

Grodner, D. and Gibson, E. (2005). Some consequences of the serial nature of linguistic input. *Cognitive Science*, 29(2):261–290.

Traxler, M. J., Morris, R. K., and Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 47:69–90.