Probabilistic Methods in Linguistics
Lecture 14: Logistic regression

Roger Levy

UC San Diego
Department of Linguistics

November 15, 2012
What we’ve covered so far in GLMs

- The form of the generalized linear model
- The special case of linear regression
- Matrix representation of linear regression
- The maximum-likelihood estimate of model parameters $\hat{\beta}$
- An unbiased estimate $s^2$ for the error variance
- Frequentist confidence regions for linear regression
- Confidence intervals and null-hypothesis significance testing for single regression parameters, using the $t$ statistics
- Problems of credit assignment in multiple linear regression
- Partitioning of variance and basic ANOVA: Null-hypothesis significance testing with the $F$ test
- Coefficient of determination (model $R^2$)
- Dealing with categorical predictors
- Interactions among predictors
- Non-linear effects of model predictors
What we’ll cover today

- Dichotomous response variables: Logistic regression
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 (?)

\[
\begin{align*}
&\text{Sally sent } [\text{the children}]_{\text{Recip}} [\text{toys}]_{\text{Theme}} & \text{Double Object} \\
&\text{Sally sent } [\text{toys}]_{\text{Theme}} \text{ to } [\text{the children}]_{\text{Recip}} & \text{Prepositional Object}
\end{align*}
\]
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 (?)

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

\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon \]
\[ X_1 = 1 \text{ iff theme indefinite, } X_2 = 1 \text{ iff recipient indefinite (both 0 otherwise)} \]

1. **Bad predictions for individual observations:** in linear regression, the noise term \( \epsilon \) is *Gaussian* (normally distributed)—it predicts that any continuous value is possible
Problems for linear models with categorical response

\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad X_1 = 1 \iff \text{theme indefinite}, \quad X_2 = 1 \iff \text{recipient indefinite (both 0 otherwise)} \]

1. **Bad predictions for individual observations:** In linear regression, the noise term \( \epsilon \) 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

\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad X_1 = 1 \text{ iff theme indefinite, } X_2 = 1 \text{ iff recipient indefinite (both 0 otherwise)} \]

1. Bad predictions for individual observations: in linear regression, the noise term \( \epsilon \) is \textit{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!

<table>
<thead>
<tr>
<th>Definiteness of recipient and theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Realization</td>
</tr>
<tr>
<td>Rec=def, Theme=def</td>
</tr>
<tr>
<td>Rec=def, Theme=indef</td>
</tr>
<tr>
<td>Double obj.</td>
</tr>
<tr>
<td>Prep. obj.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Definiteness of recipient and theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Realization</td>
</tr>
<tr>
<td>Rec=indef, Theme=def</td>
</tr>
<tr>
<td>Rec=indef, Theme=indef</td>
</tr>
<tr>
<td>Double obj.</td>
</tr>
<tr>
<td>Prep. obj.</td>
</tr>
</tbody>
</table>
Problems for linear models with categorical response

\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad X_1 = 1 \text{ iff theme indefinite, } X_2 = 1 \text{ iff recipient indefinite (both 0 otherwise)} \]

1. **Bad predictions for individual observations:** in linear regression, the noise term \( \epsilon \) 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 |

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
  - Mary sent *John* a shiny toy
  - Mary sent *her friend* a shiny toy
  - Mary sent *every kid in the room* a shiny toy

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

- Recipient is PP
  - Mary sent a shiny toy to John
  - Mary sent a shiny toy to her friend
  - 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 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 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 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!
Assumptions of the generalized linear model (GLM):

1. Predictors $\{X_i\}$ influence $Y$ through the mediation of a linear predictor $\eta$;
Assumptions of the generalized linear model (GLM):

1. Predictors \( \{X_i\} \) influence \( Y \) through the mediation of a linear predictor \( \eta \);
2. \( \eta \) is a linear combination of the \( \{X_i\} \):
Assumptions of the generalized linear model (GLM):

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

\[
\eta = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \quad \text{(linear predictor)}
\]
Reviewing GLMs

Assumptions of the generalized linear model (GLM):

1. Predictors \( \{X_i\} \) influence \( Y \) through the mediation of a linear predictor \( \eta \);
2. \( \eta \) is a linear combination of the \( \{X_i\} \):
   \[
   \eta = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \quad \text{(linear predictor)}
   \]
3. \( \eta \) determines the predicted mean \( \mu \) of \( Y \)
   \[
   \eta = l(\mu) \quad \text{(link function)}
   \]
Reviewing GLMs

Assumptions of the generalized linear model (GLM):

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

2. $\eta$ is a linear combination of the $\{X_i\}$:

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

3. $\eta$ determines the predicted mean $\mu$ of $Y$

$$\eta = \ell(\mu) \quad \text{(link function)}$$

4. There is some noise distribution of $Y$ around the predicted mean $\mu$ 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 \( \mu \):
  \[ P(Y = y | \mu) = \begin{cases} 
\mu & y = 1 \\
1 - \mu & y = 0 \\
0 & \text{otherwise}
\end{cases} \]
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 \( \mu \):
  \[ P(Y = y | \mu) = \begin{cases} 
  \mu & y = 1 \\
  1 - \mu & y = 0 \\
  0 & \text{otherwise}
\end{cases} \]
- 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 \( \mu \):
  \[ P(Y = y|\mu) = \begin{cases} 
  \mu & y = 1 \\
  1 - \mu & y = 0 \\
  0 & \text{otherwise}
\end{cases} \]
- 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:

\[ \eta = \log \left( \frac{\mu}{1 - \mu} \right) \]

- 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

\[ \eta = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \]

\[ \mu = \frac{e^{\eta}}{1 + e^{\eta}} \]

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

- As with linear regression, the regression weights \( \langle \alpha, \beta_1, \ldots, \beta_m \rangle \) must be learned
Estimating parameters in logistic regression

\[ \eta = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \]

\[ \mu = \frac{e^\eta}{1 + e^\eta} \]

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

- 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 (\( \sigma^2 \) in linear regression).
Estimating parameters in logistic regression

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

- 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 (\( \sigma^2 \) in linear regression)
- Once again, we can use the method of maximum likelihood 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:

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

\[
\mu = \frac{e^\eta}{1 + e^\eta}
\]

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

The maximum likelihood estimate for the three regression parameters is

\[
\hat{\alpha} = -0.61 \\
\hat{\beta}_{\text{ThemeDef}} = 1.8 \\
\hat{\beta}_{\text{RecDef}} = -2.2
\]
Interpreting an additive logistic regression model

\[ \hat{\alpha} = -0.61 \]
\[ \hat{\beta}_{\text{ThemeDef}} = 1.8 \]
\[ \hat{\beta}_{\text{RecDef}} = -2.2 \]
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:

$$\eta = X\beta$$
Confidence regions for logistic regression

- Let us write the linear-predictor part of our GLM in matrix form:
  \[ \eta = X\beta \]

- In linear regression, we built confidence regions for parameter estimates on the basis that the covariance matrix of the MLE \( \hat{\beta} \) can be written exactly as
  \[ \text{Cov}(\hat{\beta}) = \sigma^2(X^T X)^{-1} \]
Confidence regions for logistic regression

- Let us write the linear-predictor part of our GLM in matrix form:
  \[ \eta = X\beta \]

- In linear regression, we built confidence regions for parameter estimates on the basis that the covariance matrix of the MLE \( \hat{\beta} \) can be written exactly as
  \[ \text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \]

- 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:

\[ \eta = X\beta \]

In linear regression, we built confidence regions for parameter estimates on the basis that the covariance matrix of the MLE \( \hat{\beta} \) can be written exactly as

\[
\text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}
\]

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 \( \hat{\beta} \) is

\[
\text{Cov}(\hat{\beta}) = \begin{bmatrix}
\frac{\partial^2 L(\beta_1)}{\partial \beta_1^2} & \frac{\partial^2 L(\beta_2)}{\partial \beta_1 \beta_2} & \cdots & \frac{\partial^2 L(\beta_m)}{\partial \beta_1 \beta_m} \\
\frac{\partial^2 L(\beta_1)}{\partial \beta_1 \beta_2} & \frac{\partial^2 L(\beta_2)}{\partial \beta_2^2} & \cdots & \frac{\partial^2 L(\beta_m)}{\partial \beta_2 \beta_m} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 L(\beta_1)}{\partial \beta_1 \beta_m} & \frac{\partial^2 L(\beta_2)}{\partial \beta_2 \beta_m} & \cdots & \frac{\partial^2 L(\beta_m)}{\partial \beta_m^2}
\end{bmatrix}
\]

(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 \(\beta'\), the quantity

\[
(\hat{\beta}' - \beta)^T \left( \text{Cov}(\hat{\beta}') \right)^{-1} (\hat{\beta}' - \beta)^T
\]

(a multivariate Wald statistic) is asymptotically \(\chi^2_k\) 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 $\beta'$, the quantity

$$
(\hat{\beta}' - \beta)^T \left( \text{Cov}(\hat{\beta}') \right)^{-1} (\hat{\beta}' - \beta)^T
$$

(a multivariate Wald statistic) is asymptotically $\chi^2_k$ distributed.

- For a single model parameter $\beta$, we can equivalently say that

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

is asymptotically normally distributed, where $SE(\hat{\beta}) = \sqrt{\text{Var}(\hat{\beta})}$. 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 \( \beta' \), the quantity

\[
(\hat{\beta'} - \beta)^T \left( \text{Cov}(\hat{\beta'}) \right)^{-1} (\hat{\beta'} - \beta)^T
\]

(a multivariate Wald statistic) is asymptotically \( \chi_k^2 \) distributed.

- For a single model parameter \( \beta \), we can equivalently say that

\[
\frac{(\hat{\beta} - \beta)}{SE(\hat{\beta})}
\]

is asymptotically normally distributed, where

\[
SE(\hat{\beta}) = \sqrt{\text{Var}(\hat{\beta})}. \]

This quantity for \( \beta = 0 \) is often called the Wald z-statistic.

- Caution! These approximations break down when the estimates \( \hat{\beta} \) 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:

The correlation between $\hat{\beta}_{\text{RecDef}}$ and $\hat{\beta}_{\text{ThemeDef}}$ is $-0.18 \rightarrow$ not much of a credit-assignment problem
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:

\[
\eta = \alpha + \beta_{\text{RecDef}} X_{\text{RecDef}} + \beta_{\text{ThemeDef}} X_{\text{ThemeDef}} + \beta_{\text{RecDef:ThemeDef}} X_{\text{RecDef}} X_{\text{ThemeDef}}
\]
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:

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

- 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:

\[
\hat{\alpha} = -0.5819215 \\
\hat{\beta}_{\text{RecDef}} = -16 \\
\hat{\beta}_{\text{ThemeDef}} = 1.803136 \\
\hat{\beta}_{\text{RecDef:ThemeDef}} = 13.952
\]

- However, the standard error of \( \hat{\beta}_{\text{RecDef:ThemeDef}} \) is huge: 1151563
Interactions in logistic regression

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

  \[ \hat{\alpha} = -0.5819215 \]
  \[ \hat{\beta}_{\text{RecDef}} = -16 \]
  \[ \hat{\beta}_{\text{ThemeDef}} = 1.803136 \]
  \[ \hat{\beta}_{\text{RecDef:ThemeDef}} = 13.952 \]

- However, the standard error of \( \hat{\beta}_{\text{RecDef:ThemeDef}} \) is huge: 1151563

- This ultimately arose because there was a *perfect prediction* possible:

<table>
<thead>
<tr>
<th>Realization</th>
<th>Definiteness of recipient and theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rec=indef, Theme=def</td>
<td></td>
</tr>
<tr>
<td>Double obj.</td>
<td>0</td>
</tr>
<tr>
<td>Prep. obj.</td>
<td>5</td>
</tr>
</tbody>
</table>
Interactions in logistic regression

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

  \[ \hat{\alpha} = -0.5819215 \]
  \[ \hat{\beta}_{\text{RecDef}} = -16 \]
  \[ \hat{\beta}_{\text{ThemeDef}} = 1.803136 \]
  \[ \hat{\beta}_{\text{RecDef:ThemeDef}} = 13.952 \]

- However, the standard error of \( \hat{\beta}_{\text{RecDef:ThemeDef}} \) is huge: 1151563

- This ultimately arose because there was a *perfect prediction* possible:

<table>
<thead>
<tr>
<th>Definiteness of recipient and theme</th>
<th>Realization</th>
<th>Rec=indef, Theme=def</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double obj.</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Prep. obj.</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

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

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

is 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

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

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

  
  is 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

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

is 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 $\chi^2_1$ 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

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

is 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 $\chi^2_1$ 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

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

is 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 $\chi^2_1$ for 0.66 is 0.582, so we conclude that the interaction is statistically insignificant.
Bringing in more predictors

- For these naturally occurring data, of course, constituent definiteness is correlated with other properties (}
Bringing in more predictors

- For these naturally occurring data, of course, constituent definiteness is correlated with other properties: 

  Recipient definiteness and pronominality: 

<table>
<thead>
<tr>
<th>PronomOfRec</th>
<th>DefinOfRec</th>
<th>nonpronominal</th>
<th>pronominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>definite</td>
<td>30</td>
<td>124</td>
<td></td>
</tr>
<tr>
<td>indefinite</td>
<td>16</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
Bringing in more predictors

- For these naturally occurring data, of course, constituent definiteness is correlated with other properties:

  Recipient definiteness and pronominality: @

  \[
  \begin{array}{cccc}
  \text{PronomOfRec} & \text{DefinOfRec} & \text{nonpronominal} & \text{pronominal} \\
  \text{definite} & 30 & 124 \\
  \text{indefinite} & 16 & 2 \\
  \end{array}
  \]

  Theme definiteness and pronominality: @

  \[
  \begin{array}{cccc}
  \text{PronomOfTheme} & \text{DefinOfTheme} & \text{nonpronominal} & \text{pronominal} \\
  \text{definite} & 29 & 29 \\
  \text{indefinite} & 110 & 4 \\
  \end{array}
  \]