Regression modeling

- Thus far, we have been concerned primarily with learning and inferences about *unconditional distributions*
Thus far, we have been concerned primarily with learning and inferences about \textit{unconditional distributions}.

- \( P(X) \) where \( X \) is formant frequency, or whether a binomial satisfies a given ordering constraint.
Regression modeling

Thus far, we have been concerned primarily with learning and inferences about *unconditional distributions*

- $P(X)$ where $X$ is formant frequency, or whether a binomial satisfies a given ordering constraint
- $P(X, Y)$ where $X$ and $Y$ are whether each of two conjuncts in an NP coordination contains a PP
Regression modeling

Thus far, we have been concerned primarily with learning and inferences about \textit{unconditional distributions}.

- $P(X)$ where $X$ is formant frequency, or whether a binomial satisfies a given ordering constraint.
- $P(X, Y)$ where $X$ and $Y$ are whether each of two conjuncts in an NP coordination contains a PP.

We have used graphical models to place constraints on the form of a joint distribution $P(X_1, \ldots, X_n)$.
Regression modeling

- However, there are many cases where we want to learn and draw inferences about *conditional distributions* \( P(Y|X_1, \ldots, X_n) \)
Regression modeling

- However, there are many cases where we want to learn and draw inferences about *conditional distributions*

\[ P(Y|X_1, \ldots, X_n) \]
Regression modeling

- However, there are many cases where we want to learn and draw inferences about conditional distributions

\[ P(Y|X_1, \ldots, X_n) \]

(Note that there may or may not be connections among \( X_1, \ldots, X_n \))
Regression modeling

- However, there are many cases where we want to learn and draw inferences about *conditional distributions* $P(Y|X_1, \ldots, X_n)$

  \[
  P(Y|X_1, \ldots, X_n)
  \]

- (Note that there may or may not be connections among $X_1, \ldots, X_n$)
- Questions one might ask:
Regression modeling

- However, there are many cases where we want to learn and draw inferences about conditional distributions

\[ P(Y|X_1, \ldots, X_n) \]

- (Note that there may or may not be connections among \(X_1, \ldots, X_n\))
- Questions one might ask:
  - Is there evidence that each \(X_i\) influences \(Y\) above and beyond the influence of the other \(X_i\)?
Regression modeling

- However, there are many cases where we want to learn and draw inferences about *conditional distributions*

\[ P(Y|X_1, \ldots, X_n) \]

- (Note that there may or may not be connections among \( X_1, \ldots, X_n \))

- Questions one might ask:
  - Is there evidence that each \( X_i \) influences \( Y \) above and beyond the influence of the other \( X_i \)?
  - Do \( X_i \) and \( X_j \) have “independent” influences on \( Y \), or do they “interact” in their influence on \( Y \)?
Regression modeling

- However, there are many cases where we want to learn and draw inferences about *conditional distributions*

\[ P(Y|X_1, \ldots, X_n) \]

- (Note that there may or may not be connections among \(X_1, \ldots, X_n\))
- Questions one might ask:
  - Is there evidence that each \(X_i\) influences \(Y\) above and beyond the influence of the other \(X_i\)?
  - Do \(X_i\) and \(X_j\) have “independent” influences on \(Y\), or do they “interact” in their influence on \(Y\)?
  - What is the *shape* of the relationship between the \(X\)’s and \(Y\)?
Example: the dative alternation (Bresnan et al., 2007)

Sally gave [the children]_{Recip} [toys]_{Theme}
Sally gave [toys]_{Theme} to [the children]_{Recip}

Double Object
Prepositional Object

- Which construction is used to express this outcome is a (more or less!) dichotomous outcome
Example: the dative alternation (Bresnan et al., 2007)

Sally gave [the children]_{Recip} [toys]_{Theme}  \quad \text{Double Object}
Sally gave [toys]_{Theme} to [the children]_{Recip}  \quad \text{Prepositional Object}

- Which construction is used to express this outcome is a (more or less!) dichotomous outcome
- Probability of PO outcome can be modeled with Bernoulli distribution, success parameter $\pi$
Example: the dative alternation (Bresnan et al., 2007)

Sally gave [the children]$_{\text{Recip}}$ [toys]$_{\text{Theme}}$  \hspace{2cm} \text{Double Object}
Sally gave [toys]$_{\text{Theme}}$ to [the children]$_{\text{Recip}}$  \hspace{2cm} \text{Prepositional Object}

- Which construction is used to express this outcome is a (more or less!) dichotomous outcome
- $\rightarrow$ Probability of PO outcome can be modeled with Bernoulli distribution, success parameter $\pi$
- There are many variables $X_i$ that may influence speaker preference
Example: the dative alternation (Bresnan et al., 2007)

Sally gave [the children]$_{\text{Recip}}$ [toys]$_{\text{Theme}}$
Sally gave [toys]$_{\text{Theme}}$ to [the children]$_{\text{Recip}}$

- Which construction is used to express this outcome is a (more or less!) dichotomous outcome
- Probability of PO outcome can be modeled with Bernoulli distribution, success parameter $\pi$
- There are many variables $X_i$ that may influence speaker preference
  - Definiteness of theme
Example: the dative alternation (Bresnan et al., 2007)

Sally gave [the children][\text{Recip}] [toys][\text{Theme}]  
Sally gave [toys][\text{Theme}] to [the children][\text{Recip}]

- Which construction is used to express this outcome is a (more or less!) dichotomous outcome
- Probability of PO outcome can be modeled with Bernoulli distribution, success parameter $\pi$
- There are many variables $X_i$ that may influence speaker preference
  - Definiteness of theme
  - Definiteness of recipient
Example: the dative alternation (Bresnan et al., 2007)

Sally gave \([\text{the children}]_{\text{Recip}} [\text{toys}]_{\text{Theme}}\)  
Sally gave \([\text{toys}]_{\text{Theme}}\) to \([\text{the children}]_{\text{Recip}}\)

- Which construction is used to express this outcome is a (more or less!) dichotomous outcome
- Probability of PO outcome can be modeled with Bernoulli distribution, success parameter \(\pi\)
- There are many variables \(X_i\) that may influence speaker preference
  - Definiteness of theme
  - Definiteness of recipient
  - Size of theme
Example: the dative alternation (Bresnan et al., 2007)

Sally gave [the children] \text{Recip} [toys] \text{Theme} \quad \text{Double Object}
Sally gave [toys] \text{Theme} to [the children] \text{Recip} \quad \text{Prepositional Object}

Which construction is used to express this outcome is a (more or less!) dichotomous outcome

→ Probability of PO outcome can be modeled with Bernoulli distribution, success parameter \( \pi \)

There are many variables \( X_i \) that may influence speaker preference

- Definiteness of theme
- Definiteness of recipient
- Size of theme
- Size of recipient
Example: the dative alternation (Bresnan et al., 2007)

Sally gave \([\text{the children}]_{\text{Recip}} [\text{toys}]_{\text{Theme}}\)  
Sally gave \([\text{toys}]_{\text{Theme}} \text{to} [\text{the children}]_{\text{Recip}}\)

- Which construction is used to express this outcome is a (more or less!) dichotomous outcome
- Probability of PO outcome can be modeled with Bernoulli distribution, success parameter \(\pi\)
- There are many variables \(X_i\) that may influence speaker preference
  - Definiteness of theme
  - Definiteness of recipient
  - Size of theme
  - Size of recipient
  - ...
Example: the dative alternation (Bresnan et al., 2007)

Sally gave [the children]_{Recip} [toys]_{Theme}  \hspace{1cm} \textbf{Double Object}
Sally gave [toys]_{Theme} to [the children]_{Recip}  \hspace{1cm} \textbf{Prepositional Object}

- Which construction is used to express this outcome is a (more or less!) dichotomous outcome
- \rightarrow Probability of PO outcome can be modeled with Bernoulli distribution, success parameter $\pi$
- There are many variables $X_i$ that may influence speaker preference
  - Definiteness of theme
  - Definiteness of recipient
  - Size of theme
  - Size of recipient
  - \ldots
- Each unique combination of these $\{X_i\}$ may potentially have its own unique value of $\pi$
Learning conditional distributions

**Example:** definiteness of recipient and definiteness of theme

Contingency table from Bresnan et al. (2007):

<table>
<thead>
<tr>
<th>Realization</th>
<th>Rec=def, Theme=def</th>
<th>Rec=def, Theme=indef</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double obj.</td>
<td>456</td>
<td>1754</td>
</tr>
<tr>
<td>Prep. obj.</td>
<td>317</td>
<td>248</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Realization</th>
<th>Rec=indef, Theme=def</th>
<th>Rec=indef, Theme=indef</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double obj.</td>
<td>34</td>
<td>170</td>
</tr>
<tr>
<td>Prep. obj.</td>
<td>122</td>
<td>162</td>
</tr>
</tbody>
</table>
Learning conditional distributions

**Example:** definiteness of recipient and definiteness of theme

Contingency table from Bresnan et al. (2007):

<table>
<thead>
<tr>
<th>Realization</th>
<th>Rec=def, Theme=def</th>
<th>Rec=def, Theme=indef</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double obj.</td>
<td>456</td>
<td>1754</td>
</tr>
<tr>
<td>Prep. obj.</td>
<td>317</td>
<td>248</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Realization</th>
<th>Rec=indef, Theme=def</th>
<th>Rec=indef, Theme=indef</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double obj.</td>
<td>34</td>
<td>170</td>
</tr>
<tr>
<td>Prep. obj.</td>
<td>122</td>
<td>162</td>
</tr>
</tbody>
</table>

We could use relative frequency estimation to learn the distributions $P(\text{Realization}|\text{RecDef})$, $P(\text{Realization}|\text{ThemeDef})$, and $P(\text{Realization}|\text{RecDef}, \text{ThemeDef})$
Learning conditional distributions

The *unconditional* distribution $P(\text{Realization})$ has one parameter:
Learning conditional distributions

\[ P(\text{Realization} | \text{RecDef}) \] has two parameters:

- Definiteness of recipient
- Probability of double object realization

![Graph showing the probability of double object realization for definities (def) and indefinitities (indef).]
Learning conditional distributions

$P(\text{Realization} | \text{ThemeDef})$ has two parameters:

- **def**
  - Definiteness of theme
  - Probability of double object realization
  - Values: 0.0, 0.2, 0.4, 0.6, 0.8, 1.0

- **indef**
  - Definiteness of theme
  - Probability of double object realization
  - Values: 0.0, 0.2, 0.4, 0.6, 0.8, 1.0
Learning conditional distributions

\[ P(\text{Realization}|\text{RecDef}, \text{ThemeDef}) \] has four parameters:

- Definiteness of recipient and theme
- Probability of double object realization

![Diagram showing the probability distribution for different combinations of definiteness of recipient and theme.](image-url)
Learning conditional distributions

- We needed one parameter for the overall mean
Learning conditional distributions

- We needed one parameter for the overall mean
- Recipient definiteness always increases double-object probability in a consistent fashion (+1 parameter)
Learning conditional distributions

- We needed one parameter for the overall mean
- Recipient definiteness always *increases* double-object probability in a consistent fashion (+1 parameter)
- Likewise, theme definiteness always *decreases* double-object probability in a consistent fashion (+1 parameter)
Learning conditional distributions

- We needed one parameter for the overall mean
- Recipient definiteness always *increases* double-object probability in a consistent fashion (+1 parameter)
- Likewise, theme definiteness always *decreases* double-object probability in a consistent fashion (+1 parameter)
- It would be nice to be able to learn a 3-parameter model that encodes these two effects
Learning conditional distributions

- We needed one parameter for the overall mean.
- Recipient definiteness always increases double-object probability in a consistent fashion (+1 parameter).
- Likewise, theme definiteness always decreases double-object probability in a consistent fashion (+1 parameter).
- It would be nice to be able to learn a 3-parameter model that encodes these two effects.
- But relative frequency estimation doesn’t give us the tools to do this!
Learning conditional distributions

- We needed one parameter for the overall mean
- Recipient definiteness always *increases* double-object probability in a consistent fashion (+1 parameter)
- Likewise, theme definiteness always *decreases* double-object probability in a consistent fashion (+1 parameter)
- It would be nice to be able to learn a 3-parameter model that encodes these two effects
- But relative frequency estimation doesn’t give us the tools to do this!
  - RFE only gives us the tools for a separate multinomial per combination of $X_i$

\[
P(\text{Realization}=\text{DO}|\text{Rec}=\text{def}, \text{Theme}=\text{def}) = \pi_1
\]
\[
P(\text{Realization}=\text{DO}|\text{Rec}=\text{def}, \text{Theme}=\text{indef}) = \pi_2
\]
\[
P(\text{Realization}=\text{DO}|\text{Rec}=\text{indef}, \text{Theme}=\text{def}) = \pi_3
\]
\[
P(\text{Realization}=\text{DO}|\text{Rec}=\text{indef}, \text{Theme}=\text{indef}) = \pi_4
\]
Learning distributions conditional on continuous RVs

- Another example: the relationship between word frequency $X$ and lexical decision time $Y$
Learning distributions conditional on continuous RVs

- Another example: the relationship between word frequency $X$ and lexical decision time $Y$

- There is no hope of using any method to learn an arbitrary, different distribution of $P(Y|X)$ for each different value of $X$
Learning distributions conditional on continuous RVs

- Another example: the relationship between word frequency $X$ and lexical decision time $Y$

There is no hope of using any method to learn an arbitrary, different distribution of $P(Y|X)$ for each different value of $X$

Furthermore, there is a clear, *systematic* relationship between $X$ and $Y$—we want to exploit it!
Generalized linear models I

Goal: model the effects of predictors (\textit{independent variables}) $X$ on a response (\textit{dependent variable}) $Y$. 
Generalized linear models I

Goal: model the effects of predictors (independent variables) $X$ on a response (dependent variable) $Y$.

The picture:
Generalized linear models I

Goal: model the effects of predictors (independent variables) $X$ on a response (dependent variable) $Y$.

The picture:
Generalized linear models I

Goal: model the effects of predictors (independent variables) $X$ on a response (dependent variable) $Y$.

The picture:
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 \( \{X_i\} \):

\[
\eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \quad \text{(linear predictor)}
\]
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 \( \{X_i\} \):
   \[
   \eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \quad \text{(linear predictor)}
   \]
3. \( \eta \) determines the predicted mean \( \mu \) of \( Y \)
   \[
   \eta = l(\mu) \quad \text{(link function)}
   \]
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 \( \{X_i\} \):
   \[
   \eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \quad \text{(linear predictor)}
   \]
3. \( \eta \) determines the predicted mean \( \mu \) of \( Y \)
   \[
   \eta = l(\mu) \quad \text{(link function)}
   \]
4. There is some noise distribution of \( Y \) around the predicted mean \( \mu \) of \( Y \):
   \[
P(Y = y; \mu)
   \]
GLMs III

Linear regression, which underlies ANOVA, is a kind of generalized linear model.
Linear regression, which underlies ANOVA, is a kind of generalized linear model.

- The predicted mean is just the linear predictor:

\[ \eta = I(\mu) = \mu \]
GLMs III

Linear regression, which underlies ANOVA, is a kind of generalized linear model.

- The predicted mean is just the linear predictor:

  \[ \eta = l(\mu) = \mu \]

- Noise is normally (=Gaussian) distributed around 0 with standard deviation \( \sigma \):

  \[ \epsilon \sim \mathcal{N}(0, \sigma) \]
GLMs III

Linear regression, which underlies ANOVA, is a kind of generalized linear model.

- The predicted mean is just the linear predictor:
  \[ \eta = \mu \]

- Noise is normally (Gaussian) distributed around 0 with standard deviation \( \sigma \):
  \[ \epsilon \sim \mathcal{N}(0, \sigma) \]

- This gives us the traditional linear regression equation:
  \[ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon \]
How do we fit the parameters $\beta_i$ and $\sigma$ (choose *model coefficients*)?

There are two major approaches (deeply related, yet different) in widespread use:
How do we fit the parameters $\beta_i$ and $\sigma$ (choose model coefficients)?

There are two major approaches (deeply related, yet different) in widespread use:

- The principle of maximum likelihood: pick parameter values that maximize the probability of your data $Y$

  \[
  \text{choose } \{\beta_i\} \text{ and } \sigma \text{ that make the likelihood } P(Y|\{\beta_i\}, \sigma) \text{ as large as possible}
  \]
How do we fit the parameters $\beta_i$ and $\sigma$ (choose model coefficients)?

There are two major approaches (deeply related, yet different) in widespread use:

- The principle of maximum likelihood: pick parameter values that maximize the probability of your data $Y$

  \[ \text{choose } \{\beta_i\} \text{ and } \sigma \text{ that make the likelihood } P(Y|\{\beta_i\},\sigma) \text{ as large as possible} \]

- Bayesian inference: put a probability distribution on the model parameters and update it on the basis of what parameters best explain the data
How do we fit the parameters $\beta_i$ and $\sigma$ (choose model coefficients)?

There are two major approaches (deeply related, yet different) in widespread use:

- The principle of maximum likelihood: pick parameter values that maximize the probability of your data $Y$

  \[
  \text{choose } \{\beta_i\} \text{ and } \sigma \text{ that make the likelihood } P(Y|\{\beta_i\}, \sigma) \text{ as large as possible}
  \]

- Bayesian inference: put a probability distribution on the model parameters and update it on the basis of what parameters best explain the data

  \[
  P(\{\beta_i\}, \sigma|Y) = \frac{P(Y|\{\beta_i\}, \sigma)P(\{\beta_i\}, \sigma)}{P(Y)}
  \]
How do we fit the parameters $\beta_i$ and $\sigma$ (choose model coefficients)?

There are two major approaches (deeply related, yet different) in widespread use:

- The principle of maximum likelihood: pick parameter values that maximize the probability of your data $Y$

  $\text{choose } \{\beta_i\} \text{ and } \sigma \text{ that make the likelihood } P(Y|\{\beta_i\}, \sigma) \text{ as large as possible}$

- Bayesian inference: put a probability distribution on the model parameters and update it on the basis of what parameters best explain the data

$$P(\{\beta_i\}, \sigma|Y) = \frac{\text{Likelihood} \cdot \text{Prior}}{P(Y)}$$
GLMs V: a simple example

- You are studying non-word RTs in a lexical-decision task
GLMs V: a simple example

- You are studying non-word RTs in a lexical-decision task
  tpozt  Word or non-word?
GLMs V: a simple example

- You are studying non-word RTs in a lexical-decision task
  
  tpozt  \textit{Word or non-word?}
  houze  \textit{Word or non-word?}
GLMs V: a simple example

- You are studying non-word RTs in a lexical-decision task
  
  tpozt  \textit{Word or non-word?}
  houze  \textit{Word or non-word?}

- Non-words with different \textit{neighborhood densities}* should have different average RT  *(= number of neighbors of edit-distance 1)*
GLMs V: a simple example

- You are studying non-word RTs in a lexical-decision task
  - tpozt  \textit{Word or non-word?}
  - houze \textit{Word or non-word?}

- Non-words with different \textit{neighborhood densities}\(^*\) should have different average RT \(^*\)(= number of neighbors of edit-distance 1)

- A simple model: assume that neighborhood density has a \textit{linear} effect on average RT, and trial-level noise is \textit{normally distributed}\(^*\) \(^*\)(n.b. wrong–RTs are skewed—but not horrible.)
GLMs V: a simple example

- You are studying non-word RTs in a lexical-decision task
  
  tpozt \textit{Word or non-word?}
  
  houze \textit{Word or non-word?}

- Non-words with different \textit{neighborhood densities} should have different average RT \((=\text{number of neighbors of edit-distance 1})\)

- A simple model: assume that neighborhood density has a \textit{linear} effect on average RT, and trial-level noise is \textit{normally distributed} \((\text{n.b. wrong–RTs are skewed—but not horrible.})\)

- If \(x_i\) is neighborhood density, our simple model is

\[
RT_i = \alpha + \beta x_i + \epsilon_i \\
\sim N(0,\sigma)
\]
GLMs V: a simple example

▶ You are studying non-word RTs in a lexical-decision task

\[
\text{tpozt} \quad \text{Word or non-word?} \\
\text{houze} \quad \text{Word or non-word?}
\]

▶ Non-words with different neighborhood densities* should have different average RT *(= number of neighbors of edit-distance 1)*

▶ A simple model: assume that neighborhood density has a linear effect on average RT, and trial-level noise is normally distributed* *(n.b. wrong–RTs are skewed—but not horrible.)*

▶ If \( x_i \) is neighborhood density, our simple model is

\[
RT_i = \alpha + \beta x_i + \epsilon_i \sim N(0,\sigma)
\]

▶ We need to draw inferences about \( \alpha, \beta, \) and \( \sigma \)
GLMs V: a simple example

- You are studying non-word RTs in a lexical-decision task
  
  tpozt  Word or non-word?
  houze  Word or non-word?

- Non-words with different neighborhood densities* should have different average RT  *(= number of neighbors of edit-distance 1)

- A simple model: assume that neighborhood density has a linear effect on average RT, and trial-level noise is normally distributed* *(n.b. wrong–RTs are skewed—but not horrible.)

- If $x_i$ is neighborhood density, our simple model is

  $$ RT_i = \alpha + \beta x_i + \mathcal{N}(0,\sigma) $$

- We need to draw inferences about $\alpha$, $\beta$, and $\sigma$

- e.g., “Does neighborhood density affects RT?” → is $\beta$ reliably non-zero?
We’ll use length-4 nonword data from (Bicknell et al., 2010) (thanks!), such as:

*Few neighbors*  
gaty peme rixy

*Many neighbors*  
lish pait yine
We’ll use length-4 nonword data from (Bicknell et al., 2010) (thanks!), such as:

- Few neighbors: gaty, peme, rixy
- Many neighbors: lish, pait, yine

There’s a wide range of neighborhood density:
GLMs VII: maximum-likelihood model fitting

\[ RT_i = \alpha + \beta X_i + \epsilon_i \sim N(0, \sigma) \]

- Here’s a translation of our simple model into R:
  
  \[ RT \sim 1 + x \]
Here’s a translation of our simple model into R:

\[ RT \sim 1 + x \]

The noise is implicit in asking R to fit a \textit{linear} model.
Here’s a translation of our simple model into R:

\[ RT \sim 1 + x \]

The noise is implicit in asking R to fit a *linear* model.

(We can omit the 1; R assumes it unless otherwise directed.)
GLMs VII: maximum-likelihood model fitting

\[ RT_i = \alpha + \beta X_i + \epsilon_i \sim N(0, \sigma) \]

- Here’s a translation of our simple model into R:
  \[ RT \sim x \]
- The noise is implicit in asking R to fit a linear model
- (We can omit the 1; R assumes it unless otherwise directed)
GLMs VII: maximum-likelihood model fitting

\[ RT_i = \alpha + \beta X_i + \varepsilon_i \sim N(0, \sigma) \]

- Here’s a translation of our simple model into R:
  \[ RT \sim x \]
- The noise is implicit in asking R to fit a linear model
- (We can omit the 1; R assumes it unless otherwise directed)
- Example of fitting via maximum likelihood: one subject from Bicknell et al. (2010)

```r
> m <- glm(RT ~ neighbors, d, family="gaussian")
> summary(m)

Gaussian noise, implicit intercept

[...]

  Estimate Std. Error t value Pr(>|t|)
(Intercept) 382.997   26.837  14.271  <2e-16  ***
neighbors   4.828    6.553   0.737   0.466

> sqrt(summary(m)[["dispersion"]])

[1] 107.2248
```
GLMs VII: maximum-likelihood model fitting

\[ RT_i = \alpha + \beta X_i + \epsilon_i \sim N(0, \sigma) \]

- Here’s a translation of our simple model into R:
  \[ RT \sim x \]
- The noise is implicit in asking R to fit a \textit{linear} model
- (We can omit the 1; R assumes it unless otherwise directed)
- Example of fitting via maximum likelihood: one subject from Bicknell et al. (2010)

```r
> m <- glm(RT ~ neighbors, d, family="gaussian")
> summary(m)

[...
  Estimate  Std. Error    t value  Pr(>|t|)
(Intercept) 382.997      26.837     14.271 <2e-16 ***
neighbors    4.828       6.553      0.737   0.466

> sqrt(summary(m)[["dispersion"]])
[1] 107.2248
```
GLMs VII: maximum-likelihood model fitting

\[ RT_i = \alpha + \beta X_i + \epsilon_i \sim N(0,\sigma) \]

- Here’s a translation of our simple model into R:
  \[ RT \sim x \]
- The noise is implicit in asking R to fit a *linear* model
- (We can omit the 1; R assumes it unless otherwise directed)
- Example of fitting via maximum likelihood: one subject from Bicknell et al. (2010)

```r
> m <- glm(RT ~ neighbors, d, family="gaussian")
> summary(m)
[...]

\( \hat{\alpha} \)

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 382.997   | 26.837  | 14.271   | <2e-16 *** |
| neighbors   | 4.828     | 6.553   | 0.737    | 0.466     |

> sqrt(summary(m)[["dispersion"]])
[1] 107.2248
GLMs VII: maximum-likelihood model fitting

\[ RT_i = \alpha + \beta X_i + \epsilon_i \sim N(0, \sigma) \]

- Here’s a translation of our simple model into R:
  \[ RT \sim x \]
- The noise is implicit in asking R to fit a *linear* model
- (We can omit the 1; R assumes it unless otherwise directed)
- Example of fitting via maximum likelihood: one subject from Bicknell et al. (2010)

```r
> m <- glm(RT ~ neighbors, d, family="gaussian")
> summary(m)

[...

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 382.997  | 26.837     | 14.271  | <2e-16   ***|
| neighbors      | 4.828    | 6.553      | 0.737   | 0.466    |

> sqrt(summary(m)[["dispersion"]])

[1] 107.2248
```

\( \hat{\alpha} \) and \( \hat{\beta} \)
GLMs VII: maximum-likelihood model fitting

\[ RT_i = \alpha + \beta X_i + \epsilon_i \sim N(0, \sigma) \]

- Here’s a translation of our simple model into R:
  \[ RT \sim x \]
- The noise is implicit in asking R to fit a \textit{linear} model
- (We can omit the 1; R assumes it unless otherwise directed)
- Example of fitting via maximum likelihood: one subject from Bicknell et al. (2010)

```r
m <- glm(RT ~ neighbors, d, family="gaussian")
> summary(m)

Call:  
glm(formula = RT ~ neighbors, family = "gaussian", data = d)

Deviance Residuals:
      Min       1Q   Median       3Q      Max
-14.4536  -3.4309  -0.1599   3.3893  17.4353

Coefficients:  
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 382.997     26.837  14.271  < 2e-16 ***
neighbors   4.828     6.553   0.737   0.466
---

Dispersion parameter for gaussian family set to 107.2248

Null deviance: 1424.6 on 317 degrees of freedom
Residual deviance: 1382.0 on 316 degrees of freedom
corrected deviance: 1382.0 on 316 degrees of freedom
AIC: 1390.0

> sqrt(summary(m)[["dispersion"]])
[1] 107.2248
```

\( \hat{\alpha}, \hat{\beta}, \hat{\sigma} \)
GLMs: maximum-likelihood fitting VIII

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>383.00</td>
</tr>
<tr>
<td>neighbors</td>
<td>4.83</td>
</tr>
<tr>
<td>$\hat{\sigma}$</td>
<td>107.22</td>
</tr>
</tbody>
</table>
GLMs: maximum-likelihood fitting VIII

<table>
<thead>
<tr>
<th>Intercept</th>
<th>383.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>neighbors</td>
<td>4.83</td>
</tr>
<tr>
<td>$\hat{\sigma}$</td>
<td>107.22</td>
</tr>
</tbody>
</table>

- Estimated coefficients are what underlies “best linear fit” plots
GLMs: maximum-likelihood fitting VIII

- Estimated coefficients are what underlies “best linear fit” plots

<table>
<thead>
<tr>
<th>Intercept</th>
<th>383.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>neighbors</td>
<td>4.83</td>
</tr>
<tr>
<td>$\hat{\sigma}$</td>
<td>107.22</td>
</tr>
</tbody>
</table>
GLMs IX: Bayesian model fitting

$P(\{\beta_i\}, \sigma | Y) = \frac{\text{Likelihood} \cdot \text{Prior}}{P(Y)}$

- Alternative to maximum-likelihood:
  Bayesian model fitting
GLMs IX: Bayesian model fitting

Alternative to maximum-likelihood: Bayesian model fitting

Simple (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable
GLMs IX: Bayesian model fitting

\[
P(\{\beta_i\}, \sigma | Y) = \underbrace{P(Y | \{\beta_i\}, \sigma)}_{\text{Likelihood}} \cdot \underbrace{P(\{\beta_i\}, \sigma)}_{\text{Prior}}
\]

- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable
- Multiply by likelihood \(\rightarrow\) posterior probability distribution over \((\alpha, \beta, \sigma)\)
GLMs IX: Bayesian model fitting

\[ P(\{\beta_i\}, \sigma | Y) = \frac{P(Y | \{\beta_i\}, \sigma) P(\{\beta_i\}, \sigma)}{P(Y)} \]

- Alternative to maximum-likelihood: Bayesian model fitting

- Simple (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable

- Multiply by likelihood \(\rightarrow\) posterior probability distribution over \((\alpha, \beta, \sigma)\)
GLMs IX: Bayesian model fitting

\[
P(\{\beta_i\}, \sigma | Y) = \frac{\text{Likelihood}}{\text{Prior}} = \frac{P(Y|\{\beta_i\}, \sigma) P(\{\beta_i\}, \sigma)}{P(Y)}
\]

- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable
- Multiply by likelihood → posterior probability distribution over \((\alpha, \beta, \sigma)\)
- Bound the region of highest posterior probability containing 95% of probability density → HPD confidence region
GLMs IX: Bayesian model fitting

\[ P(\{\beta_i\}, \sigma | Y) = \frac{\text{Likelihood} \cdot \text{Prior}}{P(Y)} \]

- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable
- Multiply by likelihood \(\rightarrow\) posterior probability distribution over \((\alpha, \beta, \sigma)\)
- Bound the region of highest posterior probability containing 95% of probability density \(\rightarrow\) HPD confidence region

![Bayesian model fitting diagram](image)
GLMs IX: Bayesian model fitting

\[ P(\{\beta_i\}, \sigma | Y) = \frac{P(Y | \{\beta_i\}, \sigma) P(\{\beta_i\}, \sigma)}{P(Y)} \]

- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable
- Multiply by likelihood \(\rightarrow\) posterior probability distribution over \((\alpha, \beta, \sigma)\)
- Bound the region of highest posterior probability containing 95% of probability density \(\rightarrow\) HPD confidence region

\[ p_{MCMC} = 0.46 \]
