Probabilistic Methods in Linguistics
Lecture 11: Introduction to Linear Regression

Roger Levy

UC San Diego
Department of Linguistics

November 3, 2012
Generalized linear models I

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

The picture:
Generalized linear models I

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

The picture:
GLMs II

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 \( \{X_i\} \):

\[
\eta = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \quad \text{(linear predictor)}
\]
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 $\{X_i\}$:
   \[ \eta = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m \]  (linear predictor)
3. $\eta$ determines the predicted mean $\mu$ of $Y$
   \[ \eta = l(\mu) \]  (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 \( \{X_i\} \):

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

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:

\[ \eta = l(\mu) = \mu \]
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 (\(\equiv\)Gaussian) distributed around 0 with standard deviation \(\sigma\):

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

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

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

- 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:
  \[
  Y = \underbrace{\alpha + \beta_1 X_1 + \cdots + \beta_m X_m}_{\text{Predicted Mean } \mu = \eta} + \underbrace{\epsilon}_{\text{Noise } \sim \mathcal{N}(0, \sigma)}
  \]
**Linear regression**

More compact representation with matrices is very useful: for \( m \) predictors and \( n \) observations,

\[
Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon
\]

More compact representation with matrices is very useful: for \( m \) predictors and \( n \) observations,

<table>
<thead>
<tr>
<th>Data vector</th>
<th>Model matrix</th>
<th>Coefficients</th>
<th>Error vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>(length ( n ))</td>
<td>(dims ( n \times (m + 1) ))</td>
<td>(length ( m + 1 ))</td>
<td>(length ( n ))</td>
</tr>
</tbody>
</table>

\[
Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \quad \beta = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}
\]
Linear regression

More compact representation with matrices is very useful: for $m$ predictors and $n$ observations,

- **Data vector** (length $n$)
  
  $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

- **Model matrix** (dims $n \times (m + 1)$)
  
  $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix}$

- **Coefficients** (length $m + 1$)
  
  $\beta = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$

- **Error vector** (length $n$)
  
  $\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$

The linear regression equation is then specified as

$$Y = X\beta + \epsilon$$
A little linear algebra

If $\mathbf{X}$ is an $L \times M$ matrix and $\mathbf{Y}$ is an $M \times N$ matrix, then $\mathbf{X}$ and $\mathbf{Y}$ can be multiplied together; the resulting matrix $\mathbf{XY}$ is an $L \times M$ matrix. If $\mathbf{Z} = \mathbf{XY}$, the $i,j$-th entry of $\mathbf{Z}$ is:

$$Z_{ij} = \sum_{k=1}^{M} X_{ik}Y_{kj}$$
A little linear algebra

\[
Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon
\]

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i, j \)-th entry of \( Z \) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}
\]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[
X\beta = \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1m} \\
1 & x_{21} & x_{22} & \cdots & x_{2m} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nm}
\end{bmatrix} \begin{bmatrix}
\alpha \\
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_m
\end{bmatrix}
\]
A little linear algebra

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i,j \)-th entry of \( Z \) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}
\]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[
X\beta = \begin{bmatrix} 1 & \alpha \\
1 & \beta_1 \\
1 & \beta_2 \\
\vdots & \vdots \\
1 & \beta_m \end{bmatrix}
\]
A little linear algebra

\[
Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon
\]

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i, j \)-th entry of \( Z \) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}
\]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[
X\beta = \begin{bmatrix}
1\alpha + x_{11}\beta_1 \\
\vdots
\end{bmatrix}
\]
A little linear algebra

\[
Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \varepsilon
\]

\[
\text{Noise} \sim N(0, \sigma)
\]

If \(X\) is an \(L \times M\) matrix and \(Y\) is an \(M \times N\) matrix, then \(X\) and \(Y\) can be multiplied together; the resulting matrix \(XY\) is an \(L \times M\) matrix. If \(Z = XY\), the \(i,j\)-th entry of \(Z\) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}
\]

Thus for our linear regression equation (note that \(M = m + 1\)):

\[
X\beta = \begin{bmatrix} 1 \alpha + x_{11} \beta_1 + x_{12} \beta_2 \\
\end{bmatrix}
\]

\[
X = \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1m} \\
1 & x_{21} & x_{22} & \cdots & x_{2m} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nm}
\end{bmatrix}
\]

\[
\beta = \begin{bmatrix}
\alpha \\
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_m
\end{bmatrix}
\]
A little linear algebra

\[
\begin{align*}
\text{Predicted Mean} & = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon \\
\text{Noise} & \sim N(0, \sigma)
\end{align*}
\]

If \( \mathbf{X} \) is an \( L \times M \) matrix and \( \mathbf{Y} \) is an \( M \times N \) matrix, then \( \mathbf{X} \) and \( \mathbf{Y} \) can be multiplied together; the resulting matrix \( \mathbf{XY} \) is an \( L \times M \) matrix. If \( \mathbf{Z} = \mathbf{XY} \), the \( i,j \)-th entry of \( \mathbf{Z} \) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}
\]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[
\mathbf{X} \beta = \begin{bmatrix}
1 \alpha + x_{11} \beta_1 + x_{12} \beta_2 + \cdots \\
\end{bmatrix}
\]
A little linear algebra

\[ Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon \]

\[ \text{Noise} \sim N(0, \sigma) \]

\[ X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \quad \beta = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} \]

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i,j \)-th entry of \( Z \) is:

\[ Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj} \]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[ X\beta = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1m}\beta_m \end{bmatrix} \]
A little linear algebra

Predicted Mean
\[ Y = \alpha + \beta_1X_1 + \cdots + \beta_mX_m + \epsilon \]

Noise \( \sim N(0, \sigma) \)

\[
\begin{align*}
X &= \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1m} \\
1 & x_{21} & x_{22} & \cdots & x_{2m} \\
& \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nm}
\end{bmatrix} \\
\beta &= \begin{bmatrix}
\alpha \\
\beta_1 \\
\vdots \\
\beta_m
\end{bmatrix}
\end{align*}
\]

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i, j \)-th entry of \( Z \) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik}Y_{kj}
\]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[
X\beta = \begin{bmatrix}
1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1m}\beta_m \\
1\alpha
\end{bmatrix}
\]
A little linear algebra

\[
Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon
\]
\[
\text{Noise} \sim N(0, \sigma)
\]

\[
X = \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1m} \\
1 & x_{21} & x_{22} & \cdots & x_{2m} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nm}
\end{bmatrix}
\quad
\beta = \begin{bmatrix}
\alpha \\
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_m
\end{bmatrix}
\]

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i,j \)-th entry of \( Z \) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}
\]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[
X\beta = \begin{bmatrix}
1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1m}\beta_m \\
1\alpha + x_{21}\beta_1 \\
\vdots
\end{bmatrix}
\]
A little linear algebra

\[ Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon \]

\[ \text{Noise} \sim N(0, \sigma) \]

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i,j \)-th entry of \( Z \) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}
\]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[
X\beta = \begin{bmatrix}
1 \alpha + x_{11} \beta_1 + x_{12} \beta_2 + \cdots + x_{1m} \beta_m \\
1 \alpha + x_{21} \beta_1 + x_{22} \beta_2
\end{bmatrix}
\]
A little linear algebra

\[
Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon
\]

\[\text{Predicted Mean} \quad \text{Noise} \sim N(0, \sigma)\]

If \(X\) is an \(L \times M\) matrix and \(Y\) is an \(M \times N\) matrix, then \(X\) and \(Y\) can be multiplied together; the resulting matrix \(XY\) is an \(L \times M\) matrix. If \(Z = XY\), the \(i, j\)-th entry of \(Z\) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}
\]

Thus for our linear regression equation (note that \(M = m + 1\)):

\[
X\beta = \begin{bmatrix}
1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1m}\beta_m \\
1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \cdots
\end{bmatrix}
\]
A little linear algebra

\[ Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon \]

\[ \text{Noise} \sim N(0, \sigma) \]

\[ x = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \]

\[ \beta = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} \]

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i, j \)-th entry of \( Z \) is:

\[ Z_{ij} = \sum_{k=1}^{M} x_{ik} y_{kj} \]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[ X\beta = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1m}\beta_m \\ 1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2m}\beta_m \end{bmatrix} \]
A little linear algebra

\[ Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon \]
\[ \text{Noise} \sim N(0, \sigma) \]

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i,j \)-th entry of \( Z \) is:

\[ Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj} \]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[
X\beta = \begin{bmatrix}
1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1m}\beta_m \\
1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2m}\beta_m \\
\vdots
\end{bmatrix}
\]
A little linear algebra

\[ Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon \]

\[ \text{Noise} \sim \mathcal{N}(0, \sigma) \]

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i, j \)-th entry of \( Z \) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}
\]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[
X\beta = \begin{bmatrix}
1 \alpha + x_{11} \beta_1 + x_{12} \beta_2 + \cdots + x_{1m} \beta_m \\
1 \alpha + x_{21} \beta_1 + x_{22} \beta_2 + \cdots + x_{2m} \beta_m \\
\vdots \\
1 \alpha
\end{bmatrix}
\]
A little linear algebra

\[ Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon \]

\[ \text{Noise} \sim N(0, \sigma) \]

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i,j \)-th entry of \( Z \) is:

\[ Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj} \]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[ X\beta = \begin{bmatrix} 1 & x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1m}\beta_m \\ 1 & x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2m}\beta_m \\ \vdots \\ 1 & x_{n1}\beta_1 \end{bmatrix} \]
A little linear algebra

[Box: Predicted Mean: $Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon$
Noise $\sim N(0, \sigma)$]

If $X$ is an $L \times M$ matrix and $Y$ is an $M \times N$ matrix, then $X$ and $Y$ can be multiplied together; the resulting matrix $XY$ is an $L \times M$ matrix. If $Z = XY$, the $i, j$-th entry of $Z$ is:

$$Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}$$

Thus for our linear regression equation (note that $M = m + 1$):

$$X\beta = \begin{bmatrix} 1 \alpha + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1m}\beta_m \\ 1 \alpha + x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2m}\beta_m \\ \vdots \\ 1 \alpha + x_{n1}\beta_1 + x_{n2}\beta_2 \end{bmatrix}$$
A little linear algebra

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i,j \)-th entry of \( Z \) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}
\]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[
X \beta = \begin{bmatrix} 1 \alpha + x_{11} \beta_1 + x_{12} \beta_2 + \ldots + x_{1m} \beta_m \\ 1 \alpha + x_{21} \beta_1 + x_{22} \beta_2 + \ldots + x_{2m} \beta_m \\ \vdots \\ 1 \alpha + x_{n1} \beta_1 + x_{n2} \beta_2 + \ldots \end{bmatrix}
\]
A little linear algebra

\[
Y = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon
\]

Noise \sim N(0, \sigma)

If \( X \) is an \( L \times M \) matrix and \( Y \) is an \( M \times N \) matrix, then \( X \) and \( Y \) can be multiplied together; the resulting matrix \( XY \) is an \( L \times M \) matrix. If \( Z = XY \), the \( i,j \)-th entry of \( Z \) is:

\[
Z_{ij} = \sum_{k=1}^{M} X_{ik} Y_{kj}
\]

Thus for our linear regression equation (note that \( M = m + 1 \)):

\[
X\beta = \begin{bmatrix}
1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1m}\beta_m \\
1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2m}\beta_m \\
\vdots \\
1\alpha + x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{nm}\beta_m
\end{bmatrix}
\]
Linear regression

- So we have our regression equation

\[ Y = X\beta + \epsilon \]
Linear regression

- So we have our regression equation

\[ Y = X\beta + \epsilon \]
Linear regression

- So we have our regression equation

\[ Y = X\beta + \epsilon \]

- In everything we cover here, we will assume that the errors are independent: \( \epsilon_i \perp \epsilon_j \mid X, \beta \) (though some parts of linear regression hold even when these assumptions are relaxed)
Linear regression

- So we have our regression equation
  \[ Y = X\beta + \epsilon \]

- In everything we cover here, we will assume that the errors are independent: \( \epsilon_i \perp \epsilon_j \mid X, \beta \) (though some parts of linear regression hold even when these assumptions are relaxed)

- The maximum-likelihood estimate \( \hat{\beta} \) turns out to be
  \[ \hat{\beta} = (X^TX)^{-1}X^TY \]
An example

- The non-word lexical decision data of ?:
An example

- The non-word lexical decision data of ?:
An example

- The non-word lexical decision data of ?:
An example

- The non-word lexical decision data of ?: 

![Graph showing the relationship between RT (ms) and Neighbors.](image)

- The linear regression equation:

\[ RT = \alpha + \beta X + \epsilon \]

where \( X \) is \# of neighbors of the nonword being recognized
An example

- The non-word lexical decision data of question mark:

![Scatter plot showing RT (ms) vs Neighbors]

- The linear regression equation:

\[ RT = \alpha + \beta X + \epsilon \]

where \( X \) is the number of neighbors of the nonword being recognized

- The MLE parameter estimates are \( \hat{\alpha} = 383, \hat{\beta} = 4.83 \)
An example

- The non-word lexical decision data of ?:

![Graph showing the relationship between RT (milliseconds) and Neighbors.]

- The linear regression equation:

\[ RT = \alpha + \beta X + \epsilon \]

where \( X \) is \# of neighbors of the nonword being recognized

- The MLE parameter estimates are \( \hat{\alpha} = 383, \hat{\beta} = 4.83 \)
An example

An extremely important quantity in linear regression is the residual sum of squares. Define the predicted value of each sum of squares

\[ \hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_m x_{im} \]
An example

- An extremely important quantity in linear regression is the **residual sum of squares**. Define the predicted value of each sum of squares

\[ \hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_m x_{im} \]

- Then the residual sum of squares is defined as

\[ \text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]
An example

- An extremely important quantity in linear regression is the **residual sum of squares**. Define the predicted value of each sum of squares

\[ \hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_m x_{im} \]

- Then the residual sum of squares is defined as

\[ RSS = \sum_{i=1}^{n} (y - \hat{y})^2 \]

- The quantity \( s^2 = RSS/(n - m - 1) \) is an unbiased estimator of the the error variance \( \sigma^2 \)
Frequentist confidence regions for linear regression

The MLE parameter values $\hat{\beta}$ are distributed multivariate normally:

$$\hat{\beta} \sim N \left( \beta, \sigma^2 (X^T X)^{-1} \right)$$
Frequentist confidence regions for linear regression

\[ Y = X\beta + \epsilon \]

- The MLE parameter values \( \hat{\beta} \) are distributed multivariate normally:

\[ \hat{\beta} \sim N \left( \beta, \sigma^2 (X^T X)^{-1} \right) \]

- Note that, in general, *the estimates of the coefficients are correlated with one another!*
Frequentist confidence regions for linear regression

\[ Y = X\beta + \epsilon \]

- The MLE parameter values \( \hat{\beta} \) are distributed multivariate normally:

\[ \hat{\beta} \sim N \left( \beta, \sigma^2 (X^T X)^{-1} \right) \]

- Note that, in general, the estimates of the coefficients are correlated with one another!

- In our example,

\[ (X^T X)^{-1} = \begin{bmatrix} 0.06265 & -0.01186 \\ -0.01186 & 0.003734 \end{bmatrix} \]

hence the correlation between \( \hat{\alpha} \) and \( \hat{\beta} \) is -0.78
Recall that a $1 - p$ frequentist confidence interval $l$ for a parameter $\theta$ is one that, if the same procedure is used to construct intervals from many different randomly generated datasets, contain $\theta$ with probability $1 - p$. 
Frequentist confidence regions for linear regression

- For linear regression, we almost always want to estimate more than one parameter (at least an intercept and one slope)
Frequentist confidence regions for linear regression

- For linear regression, we almost always want to estimate more than one parameter (at least an intercept and one slope)
- Generalizing on confidence intervals: a $1 - p$ confidence region on a parameter vector $\theta$ will contain $\theta$ with probability $1 - p$
Frequentist confidence regions for linear regression

- For linear regression, we almost always want to estimate more than one parameter (at least an intercept and one slope).
- Generalizing on confidence intervals: a $1 - p$ confidence region on a parameter vector $\theta$ will contain $\theta$ with probability $1 - p$.
- In linear regression, we can form a confidence region on any size-$k$ subset of model parameters $\beta' \in \beta$ as follows:
Frequentist confidence regions for linear regression

- For linear regression, we almost always want to estimate more than one parameter (at least an intercept and one slope).
- Generalizing on confidence intervals: a $1 - p$ confidence region on a parameter vector $\theta$ will contain $\theta$ with probability $1 - p$.
- In linear regression, we can form a confidence region on any size-$k$ subset of model parameters $\beta' \in \beta$ as follows:
  - The quantity
    \[
    \frac{(\hat{\beta}' - \beta')^T X^T X (\hat{\beta}' - \beta)}{k s^2}
    \]
    is $F$-distributed with $k, n - m - 1$ degrees of freedom. The confidence region will always be an ellipse.
Frequentist confidence regions for linear regression

- For linear regression, we almost always want to estimate more than one parameter (at least an intercept and one slope).
- Generalizing on confidence intervals: a \(1 - p\) confidence region on a parameter vector \(\theta\) will contain \(\theta\) with probability \(1 - p\).
- In linear regression, we can form a confidence region on any size-\(k\) subset of model parameters \(\beta' \in \beta\) as follows:
  - The quantity
    \[
    \frac{(\hat{\beta}' - \beta')^T X^T X (\hat{\beta}' - \beta)}{k s^2}
    \] is \(F\)-distributed with \(k, n - m - 1\) degrees of freedom. The confidence region will always be an ellipse.
  - Suppose that \(Q_{F_{k,n-m-1}}\) is the quantile function for \(F_{k,n-m-1}\). Then the following is a \(1 - p\) confidence region:
    \[
    \frac{(\hat{\beta}' - \beta')^T X^T X (\hat{\beta}' - \beta)}{k s^2} \leq Q_{F_{k,n-m-1}}(1 - p)
    \]
Frequentist confidence regions for linear regression

- For linear regression, we almost always want to estimate more than one parameter (at least an intercept and one slope).
- Generalizing on confidence intervals: a $1 - p$ confidence region on a parameter vector $\theta$ will contain $\theta$ with probability $1 - p$.
- In linear regression, we can form a confidence region on any size-$k$ subset of model parameters $\beta' \in \beta$ as follows:
  - The quantity
    \[
    \frac{(\hat{\beta}' - \beta')^T X^T X (\hat{\beta}' - \beta)}{k s^2}
    \]
  is $F$-distributed with $k, n - m - 1$ degrees of freedom. The confidence region will always be an ellipse.
  - Suppose that $Q_{F_{k,n−m−1}}$ is the quantile function for $F_{k,n−m−1}$. Then the following is a $1 - p$ confidence region:
    \[
    \frac{(\hat{\beta}' - \beta')^T X^T X (\hat{\beta}' - \beta)}{k s^2} \leq Q_{F_{k,n−m−1}}(1 - p)
    \]
  - It will always be an ellipsoid whose shape is determined by $X^T X$ and whose size is determined by $p$ (the size of the region) and $s^2$ (the estimate of the error variance).
Frequentist confidence regions for linear regression

Our original example:
Frequentist confidence regions for linear regression

Our original example:
Frequentist confidence regions for linear regression

Our original example:
Frequentist confidence regions for linear regression

Our original example:
Frequentist confidence regions for linear regression

Our original example:
Frequentist confidence regions for linear regression

Our original example:
Null hypothesis significance testing with the $t$-statistic

- Recall: general confidence region is built on the fact that

$$
\frac{(\hat{\beta}' - \beta')^T X^T X (\hat{\beta}' - \beta)}{k s^2} \sim F_{k,n-m-1}
$$
Null hypothesis significance testing with the $t$-statistic

- Recall: general confidence region is built on the fact that
  \[
  \frac{(\hat{\beta}' - \beta')^T X^T X (\hat{\beta}' - \beta)}{k s^2} \sim F_{k,n-m-1}
  \]

- Suppose we focus on just one parameter $\beta_i$ (so $k = 1$)
Null hypothesis significance testing with the $t$-statistic

- Recall: general confidence region is built on the fact that
  \[
  \frac{(\hat{\beta}' - \beta')^T X^T X (\hat{\beta}' - \beta)}{k s^2} \sim F_{k, n-m-1}
  \]

- Suppose we focus on just one parameter $\beta_i$ (so $k = 1$)

- The general confidence region collapses down to a 1-dimensional confidence interval:
  \[
  \frac{(\hat{\beta}_i - \beta_i)(X^T X)_{ii}(\hat{\beta}_i - \beta_i)}{s^2} = (\hat{\beta}_i - \beta_i)^2 \frac{(X^T X)_{ii}}{s^2} \sim F_{1, n-m-1}
  \]

  But an $F$-distributed RV with $(1, N)$ d.f. in the numerator is the square of a $t$-distributed RV with $N$ d.f., so
  \[
  (\hat{\beta}_i - \beta_i) \frac{\sqrt{(X^T X)_{ii}}}{s} \sim t_{n-m-1}
  \]
Null hypothesis significance testing with the $t$-statistic

- Recall: general confidence region is built on the fact that
  \[ \frac{(\hat{\beta}' - \beta')^T X^T X (\hat{\beta}' - \beta)}{k s^2} \sim F_{k,n-m-1} \]

- Suppose we focus on just one parameter $\beta_i$ (so $k = 1$)

- The general confidence region collapses down to a 1-dimensional confidence interval:
  \[ \frac{(\hat{\beta}_i - \beta_i)(X^T X)_{ii}(\hat{\beta}_i - \beta_i)}{s^2} = (\hat{\beta}_i - \beta_i)^2 \frac{(X^T X)_{ii}}{s^2} \sim F_{1,n-m-1} \]

  But an $F$-distributed RV with $(1, N)$ d.f. in the numerator is the square of a $t$-distributed RV with $N$ d.f., so
  \[ (\hat{\beta}_i - \beta_i)\sqrt{(X^T X)_{ii}} \sim t_{n-m-1} \]

- The quantity $1/(X^T X)_{ii}$ is often called the standard error of the estimate $\hat{\beta}_i$
Null hypothesis significance testing with the $t$-statistic

\[(\hat{\beta}_i - \beta_i) \sqrt{ (X^T X)_{ii} \over s } \sim t_{n-m-1} \]

Suppose our null hypothesis is $H_0 : \beta_i = 0$. Then

\[\hat{\beta}_i \sqrt{(X^T X)_{ii} \over s} \]

is $t$-distributed with $n - m - 1$ degrees of freedom. This is often called the **t-value** of the parameter estimate. You can use the cumulative distribution function for the $t$ distribution to compute a significance level for rejecting the possibility that the true value of $\beta_i$ is 0.
Null hypothesis significance testing with the $t$-statistic

$$(\hat{\beta}_i - \beta_i) \frac{\sqrt{(X^TX)_{ii}}}{s} \sim t_{n-m-1}$$

- Suppose our null hypothesis is $H_0 : \beta_i = 0$. Then

$$\hat{\beta}_i \frac{\sqrt{(X^TX)_{ii}}}{s}$$

is $t$-distributed with $n - m - 1$ degrees of freedom. This is often called the $t$-value of the parameter estimate. You can use the cumulative distribution function for the $t$ distribution to compute a significance level for rejecting the possibility that the true value of $\beta_i$ is 0.

- **Example:** in our case, $\hat{\beta}_{RT} = 4.8; \ SE_{RT} = 6.6$, so the $t$-statistic of the estimate is 0.74. This is statistically insignificant