Generalized linear models, linear regression, parameter inference, the F test, credit assignment

Roger Levy

Massachusetts Institute of Technology

April 16, 2024

Regression modeling

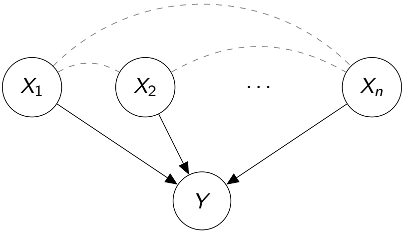
- We often want a **parameterized form** to draw inferences about *conditional distributions* $P(Y|X_1, \ldots, X_n)$

Regression modeling

- We often want a **parameterized form** to draw inferences about *conditional distributions* $P(Y|X_1, \ldots, X_n)$

Regression modeling

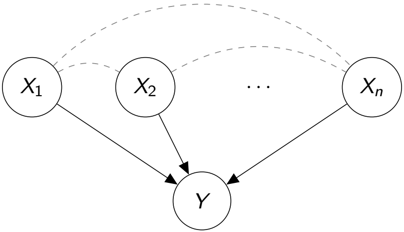
- We often want a **parameterized form** to draw inferences about *conditional distributions* $P(Y|X_1, \dots, X_n)$



- **Note!** We are setting aside for now the question of the independence/causal structure among $(X_, \dots, X_n)$

Regression modeling

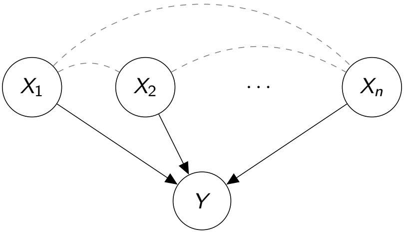
- We often want a **parameterized form** to draw inferences about *conditional distributions* $P(Y|X_1, \ldots, X_n)$



- **Note!** We are setting aside for now the question of the independence/causal structure among $(X_, \ldots, X_n)$
- Questions one might ask:

Regression modeling

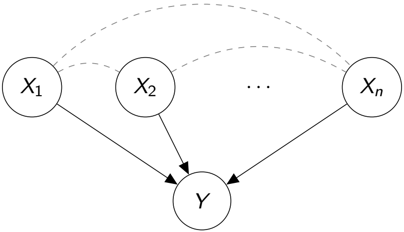
- We often want a **parameterized form** to draw inferences about *conditional distributions* $P(Y|X_1, \dots, X_n)$



- **Note!** We are setting aside for now the question of the independence/causal structure among $(X_, \dots, X_n)$
- Questions one might ask:
  - Is there evidence that each X_i predicts Y above and beyond the predictive value of the other X_i?

Regression modeling

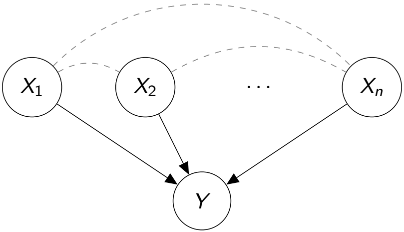
- We often want a **parameterized form** to draw inferences about *conditional distributions* $P(Y|X_1, \dots, X_n)$



- **Note!** We are setting aside for now the question of the independence/causal structure among $(X_, \dots, X_n)$
- Questions one might ask:
  - Is there evidence that each X_i predicts Y above and beyond the predictive value of the other X_i?
  - Do X_i and X_j have "separate" influences on Y, or do they "interact" in their influence on Y?

Regression modeling

- We often want a **parameterized form** to draw inferences about *conditional distributions* $P(Y|X_1,\dots,X_n)$



- **Note!** We are setting aside for now the question of the independence/causal structure among $(X_,\dots,X_n)$
- Questions one might ask:
  - Is there evidence that each X_i predicts Y above and beyond the predictive value of the other X_j?
  - Do X_i and X_j have "separate" influences on Y, or do they "interact" in their influence on Y?
  - What is the *shape* of the predictive relationship between the X's and Y?

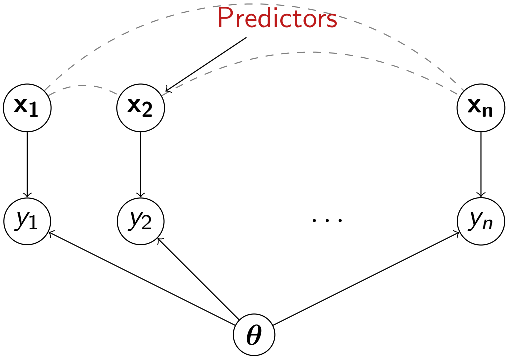
Generalized linear models I

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

Generalized linear models I

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

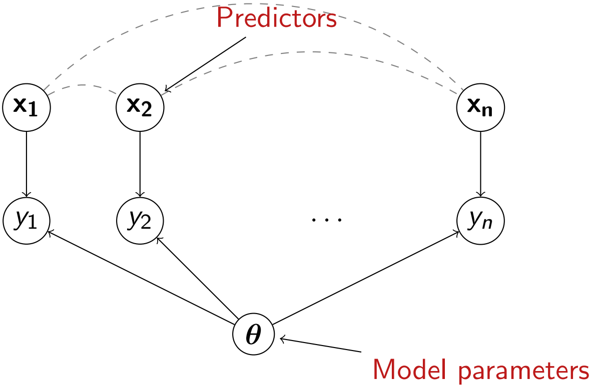
The picture:



Generalized linear models I

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

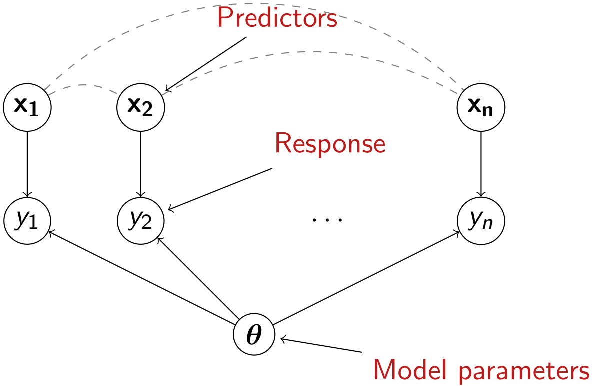
The picture:



Generalized linear models I

Goal: model the effects of predictors (independent variables) $\mathbf{X}$ 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 η;

GLMs II

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 $\{X_i\}$:

GLMs II

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 $\{X_i\}$:

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

GLMs II

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 $\{X_i\}$:

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

3. η determines the predicted mean μ of Y

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

GLMs II

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 $\{X_i\}$:

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

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:

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

GLMs III

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

GLMs III

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 σ:

$$\epsilon \sim N(0, \sigma)$$

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 σ:

$$\epsilon \sim N(0, \sigma)$$

- This gives us the traditional linear regression equation:

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean } \mu = \eta} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

GLMs IV

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

- How do we fit the parameters β_i and σ (choose *model coefficients*)?
- There are two major approaches (deeply related, yet different) in widespread use:

GLMs IV

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

- How do we fit the parameters β_i and σ (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

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

GLMs IV

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

- How do we fit the parameters β_i and σ (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

choose $\{\beta_i\}$ *and* σ *that make the likelihood* $P(Y|\{\beta_i\}, \sigma)$ *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

GLMs IV

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

- How do we fit the parameters β_i and σ (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

choose $\{\beta_i\}$ *and* σ *that make the likelihood* $P(Y|\{\beta_i\},\sigma)$ *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

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

GLMs IV

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

- How do we fit the parameters β_i and σ (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

choose $\{\beta_i\}$ *and* σ *that make the likelihood* $P(Y|\{\beta_i\},\sigma)$ *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

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

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` *Word or non-word?*
  `houze` *Word or non-word?*

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)

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.)

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

$$RT_i = \alpha + \beta x_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}$$

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

$$RT_i = \alpha + \beta x_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}$$

- We need to draw inferences about α, β, and σ

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

$$RT_i = \alpha + \beta x_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}$$

- We need to draw inferences about α, β, and σ
- e.g., "Does neighborhood density affects RT?" $\rightarrow$ is β reliably non-zero?

GLMs VI

- We'll use length-4 nonword data from (Bicknell et al., 2010) (thanks!), such as:

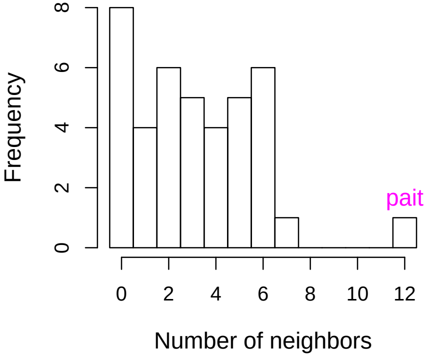
Few neighbors			*Many neighbors*		
`gaty`	`peme`	`rixy`	`lish`	`pait`	`yine`

GLMs VI

- We'll use length-4 nonword data from (Bicknell et al., 2010) (thanks!), such as:

Few neighbors			*Many neighbors*		
`gaty`	`peme`	`rixy`	`lish`	`pait`	`yine`

- There's a wide range of neighborhood density:



GLMs VII: maximum-likelihood model fitting

$$\boxed{RT_i = \alpha + \beta X_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}}$$

- Here's a translation of our simple model into R:

```
RT ~ 1 + x
```

GLMs VII: maximum-likelihood model fitting

$$\boxed{RT_i = \alpha + \beta X_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}}$$

- Here's a translation of our simple model into R:

```
RT ~ 1 + x
```

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

GLMs VII: maximum-likelihood model fitting

$$\boxed{RT_i = \alpha + \beta X_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}}$$

- Here's a translation of our simple model into `R`:

  `RT ~ 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

$$\boxed{RT_i = \alpha + \beta X_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}}$$

- Here's a translation of our simple model into `R`:

  `RT ~ 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

$$\boxed{RT_i = \alpha + \beta X_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}}$$

- Here's a translation of our simple model into R:

  RT ~ 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)

```
> 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
```

Gaussian noise, implicit intercept

GLMs VII: maximum-likelihood model fitting

$$\boxed{RT_i = \alpha + \beta X_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}}$$

- Here's a translation of our simple model into R:

  RT ~ 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)

```
> 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

$$\boxed{RT_i = \alpha + \beta X_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}}$$

- Here's a translation of our simple model into R:

$$\texttt{RT} \sim \qquad \texttt{x}$$

- 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)

```
> 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
```

$\widehat{\alpha} \longrightarrow$ Estimate, $\boxed{382.997}$

GLMs VII: maximum-likelihood model fitting

$$RT_i = \alpha + \beta X_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}$$

- 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)

```
> 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
```

$\widehat{\alpha}$ $\widehat{\beta}$

GLMs VII: maximum-likelihood model fitting

$$RT_i = \alpha + \beta X_i + \overbrace{\epsilon_i}^{\sim N(0,\sigma)}$$

- 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)

```
> 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
```

$\widehat{\alpha}$ $\widehat{\beta}$ $\widehat{\sigma}$

GLMs: maximum-likelihood fitting VIII

Intercept	383.00
neighbors	4.83
$\widehat{\sigma}$	107.22

GLMs: maximum-likelihood fitting VIII

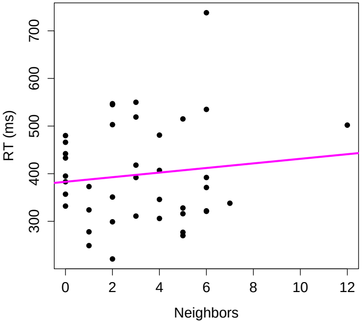
`Intercept`	`383.00`
`neighbors`	`4.83`
$\widehat{\sigma}$	`107.22`

- Estimated coefficients are what underlies "best linear fit" plots

GLMs: maximum-likelihood fitting VIII

`Intercept`	`383.00`
`neighbors`	`4.83`
$\widehat{\sigma}$	`107.22`

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



GLMs IX: Bayesian model fitting

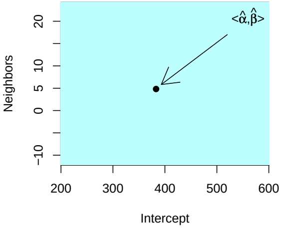
$$P(\{\beta_i\},\sigma|Y) = \frac{\overbrace{P(Y|\{\beta_i\},\sigma)}^{\text{Likelihood}}\overbrace{P(\{\beta_i\},\sigma)}^{\text{Prior}}}{P(Y)}$$

- Alternative to maximum-likelihood: Bayesian model fitting

GLMs IX: Bayesian model fitting

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

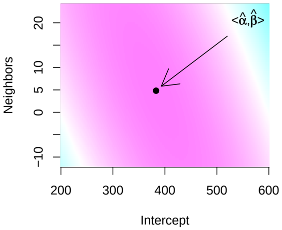
- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of (α, β, σ) equally probable



GLMs IX: Bayesian model fitting

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

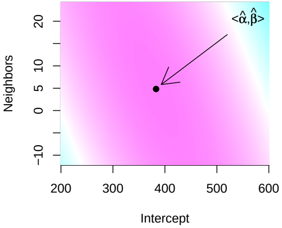
- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of (α,β,σ) equally probable
- Multiply by likelihood $\rightarrow$ posterior probability distribution over (α,β,σ)



GLMs IX: Bayesian model fitting

$$\color{red}{P(\{\beta_i\},\sigma|Y)} = \frac{\overbrace{P(Y|\{\beta_i\},\sigma)}^{\text{Likelihood}}\overbrace{P(\{\beta_i\},\sigma)}^{\text{Prior}}}{P(Y)}$$

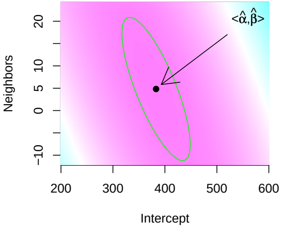
- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of (α,β,σ) equally probable
- Multiply by likelihood → posterior probability distribution over (α,β,σ)



GLMs IX: Bayesian model fitting

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

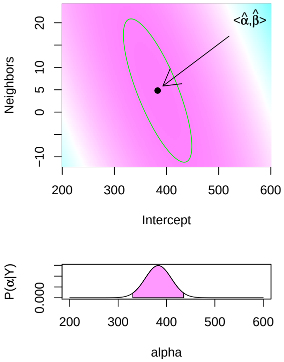
- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of (α, β, σ) equally probable
- Multiply by likelihood $\rightarrow$ posterior probability distribution over (α, β, σ)
- Bound the region of highest posterior probability containing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



GLMs IX: Bayesian model fitting

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

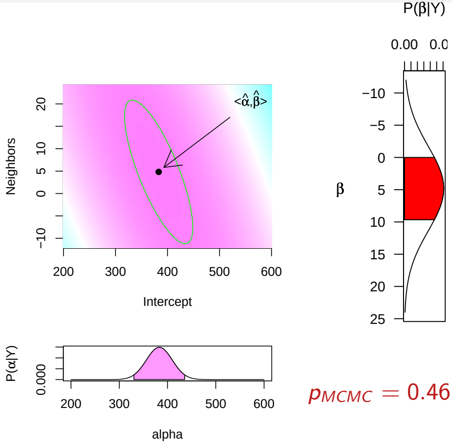
- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of (α, β, σ) equally probable
- Multiply by likelihood $\rightarrow$ posterior probability distribution over (α, β, σ)
- Bound the region of highest posterior probability containing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



GLMs IX: Bayesian model fitting

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

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



$p_{MCMC} = 0.46$

- p_{MCMC} (Baayen et al., 2008) is 1 minus the largest possible symmetric confidence interval wholly on one side of 0

Linear regression

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

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

Data vector (length n)	Model matrix (dims $n \times (m+1)$)	Coefficients (length $m+1$)	Error vector (length n)

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Linear regression

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

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

Data vector (length n)	Model matrix (dims $n \times (m+1)$)	Coefficients (length $m+1$)	Error vector (length n)
$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$	$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$	$\boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$	$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$

- The linear regression equation is then specified as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

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

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

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

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

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

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha \\ \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

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

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 \\ \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

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

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 \\ \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

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

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \ldots \\ \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m \\ \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

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

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m \\ 1\alpha \\ \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

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

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m \\ 1\alpha + x_{21}\beta_1 \\ \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1m}\beta_m \\ 1\alpha + x_{21}\beta_1 + x_{22}\beta_2 \\ \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1m}\beta_m \\ 1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \dots \\ \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m \\ 1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2m}\beta_m \\ \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m \\ 1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2m}\beta_m \\ \vdots \\ \\ \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m \\ 1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2m}\beta_m \\ \vdots \\ 1\alpha \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

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

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m \\ 1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2m}\beta_m \\ \vdots \\ 1\alpha + x_{n1}\beta_1 \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

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

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m \\ 1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2m}\beta_m \\ \vdots \\ 1\alpha + x_{n1}\beta_1 + x_{n2}\beta_2 \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

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

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 $\mathbf{Z} = \mathbf{XY}$, 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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m \\ 1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2m}\beta_m \\ \vdots \\ 1\alpha + x_{n1}\beta_1 + x_{n2}\beta_2 + \ldots \end{bmatrix}$$

A little linear algebra

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

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

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 $M = m + 1$):

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1\alpha + x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m \\ 1\alpha + x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2m}\beta_m \\ \vdots \\ 1\alpha + x_{n1}\beta_1 + x_{n2}\beta_2 + \ldots + x_{nm}\beta_m \end{bmatrix}$$

Linear regression

- So we have our regression equation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Linear regression

- So we have our regression equation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- For now, we will assume that the errors are independent given the covariates: $\epsilon_i \perp \epsilon_j \mid \mathbf{X}$ (though some parts of linear regression hold even when these assumptions are relaxed)

Linear regression

- So we have our regression equation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- For now, we will assume that the errors are independent given the covariates: $\epsilon_i \perp \epsilon_j \mid \mathbf{X}$ (though some parts of linear regression hold even when these assumptions are relaxed)
- The maximum-likelihood estimate $\widehat{\boldsymbol{\beta}}$ turns out to be

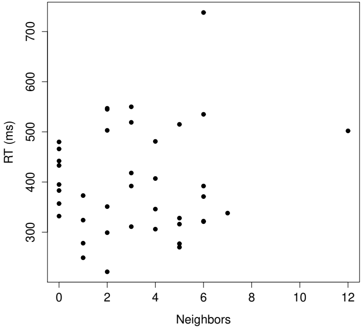
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

An example

- The non-word lexical decision data of Bicknell et al. (2010):

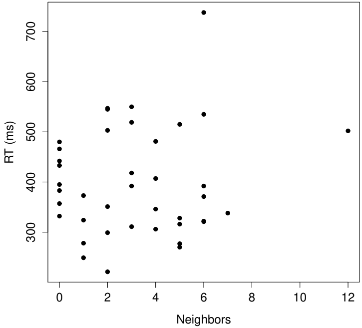
An example

- The non-word lexical decision data of Bicknell et al. (2010):



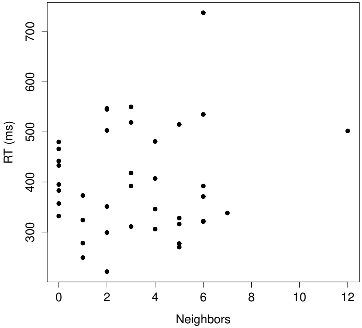
An example

- The non-word lexical decision data of Bicknell et al. (2010):



An example

- The non-word lexical decision data of Bicknell et al. (2010):



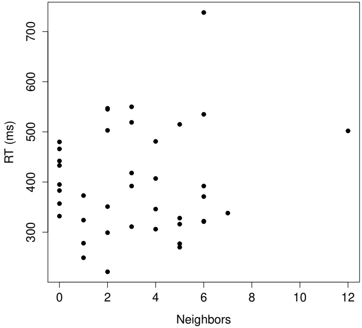
- The linear regression equation:

$$RT = \alpha + \beta X + \epsilon$$

where X is $\#$ of neighbors of the nonword being recognized

An example

- The non-word lexical decision data of Bicknell et al. (2010):



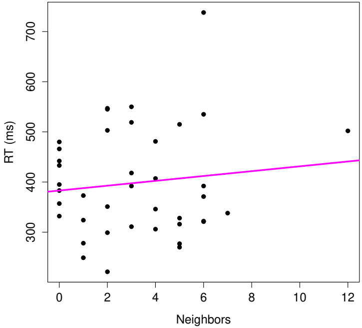
- The linear regression equation:

$$RT = \alpha + \beta X + \epsilon$$

where X is # of neighbors of the nonword being recognized
- The MLE parameter estimates are $\widehat{\alpha} = 383$, $\widehat{\beta} = 4.83$

An example

- The non-word lexical decision data of Bicknell et al. (2010):



- The linear regression equation:

$$RT = \alpha + \beta X + \epsilon$$

where X is # of neighbors of the nonword being recognized
- The MLE parameter estimates are $\widehat{\alpha} = 383$, $\widehat{\beta} = 4.83$

An example

- An key 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 key 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

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

An example

- An key 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 σ^2

Frequentist confidence regions for linear regression

$$\boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}}$$

- The MLE parameter values $\widehat{\boldsymbol{\beta}}$ are distributed multivariate normally:

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)$$

Frequentist confidence regions for linear regression

$$\boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}}$$

- The MLE parameter values $\widehat{\boldsymbol{\beta}}$ are distributed multivariate normally:

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)$$

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

Frequentist confidence regions for linear regression

$$\boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}}$$

- The MLE parameter values $\widehat{\boldsymbol{\beta}}$ are distributed multivariate normally:

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)$$

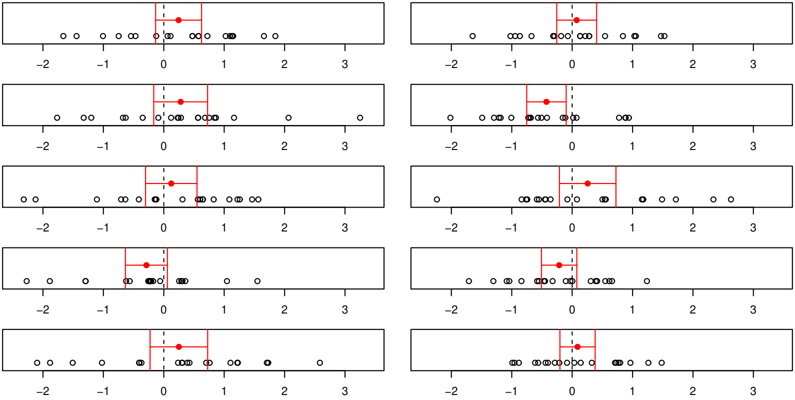
- Note that, in general, *the estimates of the coefficients are correlated with one another!*
- In our example,

$$(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \begin{bmatrix} 0.06265 & -0.01186 \\ -0.01186 & 0.003734 \end{bmatrix}$$

hence the correlation between $\widehat{\alpha}$ and $\widehat{\beta}$ is $-0.78(=\frac{-0.01186}{\sqrt{0.06265*0.003734}})$

Frequentist confidence regions for linear regression

- Recall that a $1 - p$ frequentist confidence interval I for a parameter θ is one that, if the same procedure is used to construct intervals from many different randomly generated datasets, contain θ 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 $\boldsymbol{\theta}$ will contain $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}$ will contain $\boldsymbol{\theta}$ with probability $1-p$
- In linear regression, we can form a confidence region on any size-k subset of model parameters $\boldsymbol{\beta}' \in \boldsymbol{\beta}$ 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 $\boldsymbol{\theta}$ will contain $\boldsymbol{\theta}$ with probability $1 - p$
- In linear regression, we can form a confidence region on any size-k subset of model parameters $\boldsymbol{\beta}' \in \boldsymbol{\beta}$ as follows:
  - The quantity
    $$\frac{(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^T X^T X (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})}{k\, s^2}$$
    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

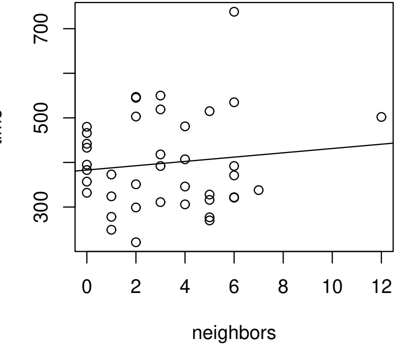
- For linear regression, we almost always want to estimate more than one parameter (at least an intercept and one slope)
- Generalizing on confidence intervals: a $1-p$ confidence region on a parameter vector $\boldsymbol{\theta}$ will contain $\boldsymbol{\theta}$ with probability $1-p$
- In linear regression, we can form a confidence region on any size-k subset of model parameters $\boldsymbol{\beta}' \in \boldsymbol{\beta}$ as follows:
  - The quantity
  $$\frac{(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^T X^T X (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})}{k\, s^2}$$
  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:
  $$\frac{(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^T X^T X (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})}{k\, s^2} \leq Q_{F_{k,n-m-1}}(1-p)$$

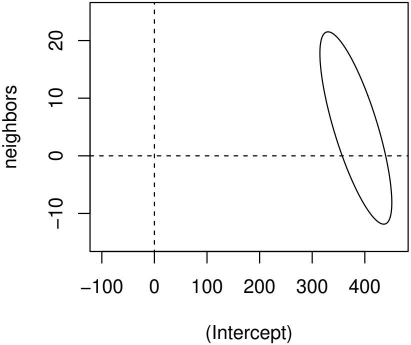
Frequentist confidence regions for linear regression

- 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 $\boldsymbol{\theta}$ will contain $\boldsymbol{\theta}$ with probability $1 - p$
- In linear regression, we can form a confidence region on any size-k subset of model parameters $\boldsymbol{\beta}' \in \boldsymbol{\beta}$ as follows:
  - The quantity
    $$\frac{(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^T X^T X (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})}{k\, s^2}$$
    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:
    $$\frac{(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^T X^T X (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})}{k\, s^2} \leq Q_{F_{k,n-m-1}}(1 - p)$$
  - 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

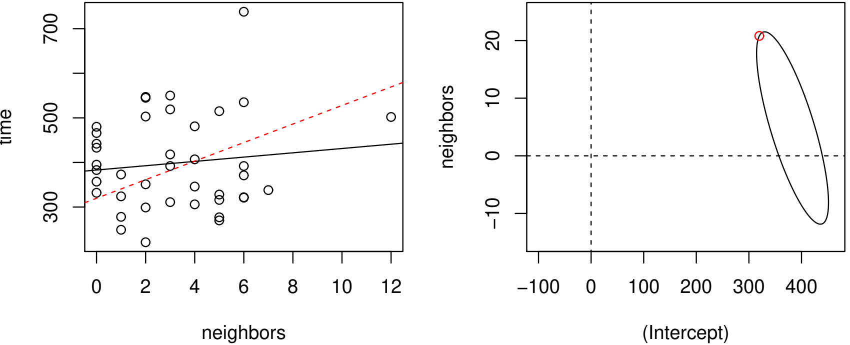
Our original example:





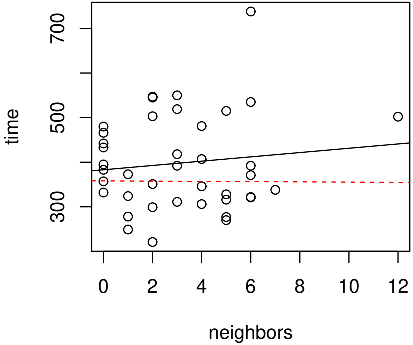
Frequentist confidence regions for linear regression

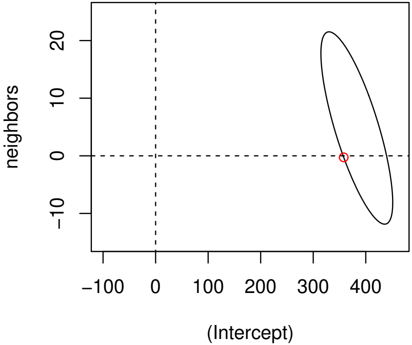
Our original example:



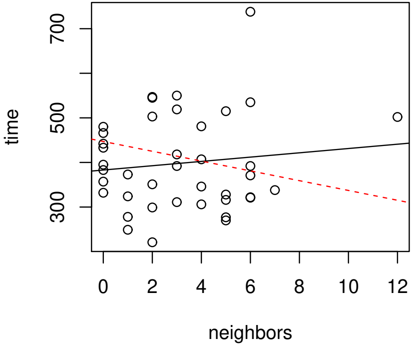
Frequentist confidence regions for linear regression

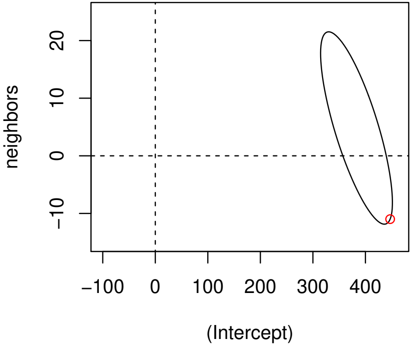
Our original example:





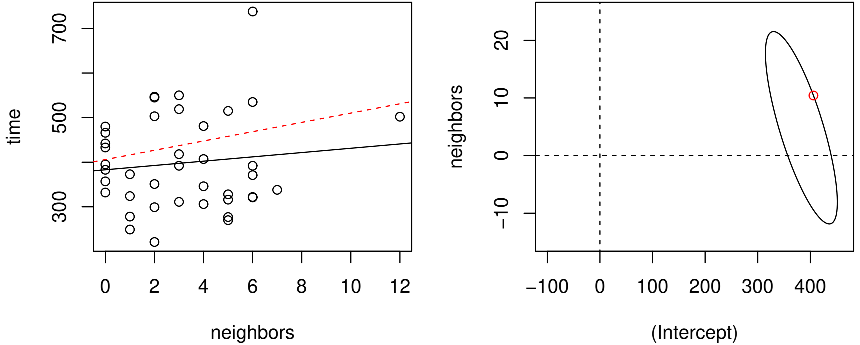
Our original example:





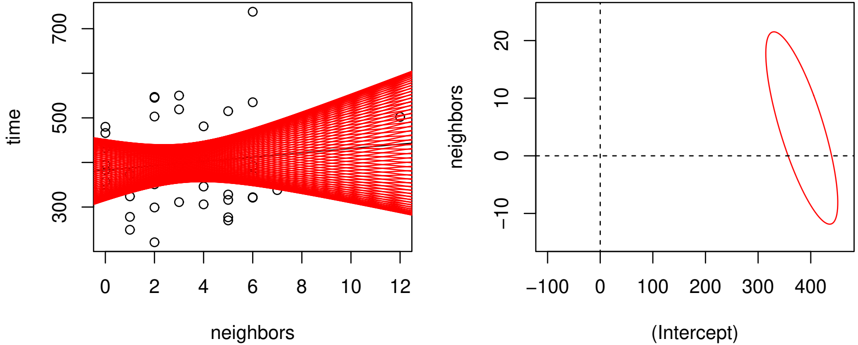
Frequentist confidence regions for linear regression

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

$$\frac{(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^T X^T X (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})}{k\, s^2} \sim F_{k,n-m-1}$$

Null hypothesis significance testing with the *t*-statistic

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

$$\frac{(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^T X^T X(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})}{k\, s^2} \sim F_{k,n-m-1}$$

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

Null hypothesis significance testing with the t-statistic

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

$$\frac{(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^T X^T X (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})}{k\, s^2} \sim F_{k,n-m-1}$$

- Suppose we focus on just one parameter β_i (so $k = 1$)
- The general confidence region collapses down to a 1-dimensional confidence interval:

$$\frac{(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)(X^T X)_{ii}(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)}{s^2} = (\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)^2 \frac{(X^T X)_{ii}}{s^2} \sim F_{1,n-m-1}$$

Null hypothesis significance testing with the t-statistic

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

$$\frac{(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^T X^T X (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})}{k\, s^2} \sim F_{k,n-m-1}$$

- Suppose we focus on just one parameter β_i (so $k = 1$)
- The general confidence region collapses down to a 1-dimensional confidence interval:

$$\frac{(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)(X^T X)_{ii}(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)}{s^2} = (\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)^2 \frac{(X^T X)_{ii}}{s^2} \sim F_{1,n-m-1}$$

- 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

$$(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) \frac{\sqrt{(X^T X)_{ii}}}{s} \sim t_{n-m-1}$$

Null hypothesis significance testing with the t-statistic

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

$$\frac{(\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^T X^T X (\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta})}{k\, s^2} \sim F_{k,n-m-1}$$

- Suppose we focus on just one parameter β_i (so $k = 1$)
- The general confidence region collapses down to a 1-dimensional confidence interval:

$$\frac{(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)(X^T X)_{ii}(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)}{s^2} = (\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)^2 \frac{(X^T X)_{ii}}{s^2} \sim F_{1,n-m-1}$$

- 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

$$(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)\frac{\sqrt{(X^T X)_{ii}}}{s} \sim t_{n-m-1}$$

- The quantity $1/\frac{(X^T X)_{ii}}{s^2}$ is often called the standard error of the estimate $\widehat{\beta}_i$

Null hypothesis significance testing with the t-statistic

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

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

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

  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 β_i is 0.

Null hypothesis significance testing with the t-statistic

$$\boxed{(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)\frac{\sqrt{(X^T X)_{ii}}}{s} \sim t_{n-m-1}}$$

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

  $$\widehat{\boldsymbol{\beta}}_i \frac{\sqrt{(X^T X)_{ii}}}{s}$$

  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 β_i is 0.
- **Example:** in our case, $\widehat{\beta}_{RT} = 4.8$; $SE_{RT} = 6.6$, so the t-statistic of the estimate is 0.74. This is statistically **insignificant**

The interpretation of regression coefficients

- Another simple psycholinguistics task: *word naming*

The interpretation of regression coefficients

- Another simple psycholinguistics task: *word naming*

The interpretation of regression coefficients

- Another simple psycholinguistics task: *word naming*

  `deed`

The interpretation of regression coefficients

- Another simple psycholinguistics task: *word naming*

  `deed`

  `share`

The interpretation of regression coefficients

- Another simple psycholinguistics task: *word naming*

  `deed`

  `share`

  `hymn`

The interpretation of regression coefficients

- Another simple psycholinguistics task: *word naming*

  ```
  deed
  share
  hymn
  stretch
  ```

- It turns out that neighborhood density also affects naming speed

The interpretation of regression coefficients

- Another simple psycholinguistics task: *word naming*

  `deed`

  `share`

  `hymn`

  `stretch`

- It turns out that neighborhood density also affects naming speed
- Additionally, we'll look at the possible predictive value of two other predictors: **word frequency** and **word length**

The interpretation of regression coefficients

```r
summary(lm(exp(RTnaming) ~ Ncount + LengthInLetters + WrittenFrequency,dat.engl
```

```
##
## Call:
## lm(formula = exp(RTnaming) ~ Ncount + LengthInLetters + WrittenFrequency,
##     data = dat.english)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.274 -13.143  -0.197  13.256  63.203
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      469.1812     9.0089  52.080  < 2e-16 ***
## Ncount            -0.7519     0.2297  -3.274 0.001130 **
## LengthInLetters    5.2222     1.4588   3.580 0.000375 ***
## WrittenFrequency  -3.4998     0.8959  -3.906 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.61 on 535 degrees of freedom
## Multiple R-squared:  0.125, Adjusted R-squared:  0.1201
## F-statistic: 25.47 on 3 and 535 DF,  p-value: 2.054e-15
```

The interpretation of regression coefficients

- The meaning of a predictor's regression coefficient is *the change in predicted response when the predictor is varied by one unit*, **if the other predictors are all held constant**

The interpretation of regression coefficients

- The meaning of a predictor's regression coefficient is *the change in predicted response when the predictor is varied by one unit*, **if the other predictors are all held constant**
- **-0.75ms per additional neighbor**

The interpretation of regression coefficients

- The meaning of a predictor's regression coefficient is *the change in predicted response when the predictor is varied by one unit*, **if the other predictors are all held constant**
- **-0.75ms per additional neighbor**
- **5.22ms per additional character of length**

The interpretation of regression coefficients

- The meaning of a predictor's regression coefficient is *the change in predicted response when the predictor is varied by one unit*, **if the other predictors are all held constant**
- **-0.75ms per additional neighbor**
- **5.22ms per additional character of length**
- **-3.50ms per doubling of word frequency**

A problem of credit assignment

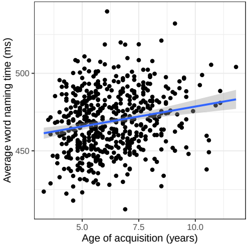
- What predicts word recognition speed: a word's *frequency* or its *age of acquisition*?

A problem of credit assignment

- What predicts word recognition speed: a word's *frequency* or its *age of acquisition*?

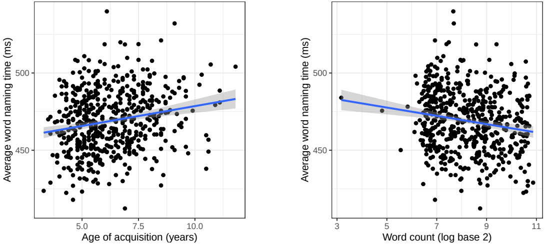
A problem of credit assignment

- What predicts word recognition speed: a word's *frequency* or its *age of acquisition*?



A problem of credit assignment

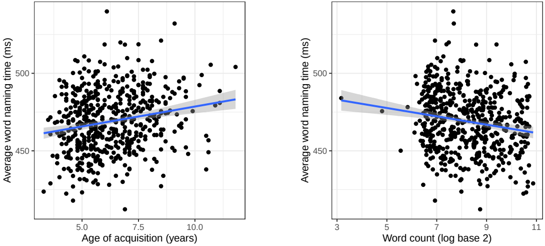
- What predicts word recognition speed: a word's *frequency* or its *age of acquisition*?



- But look at the relationship between the two!

A problem of credit assignment

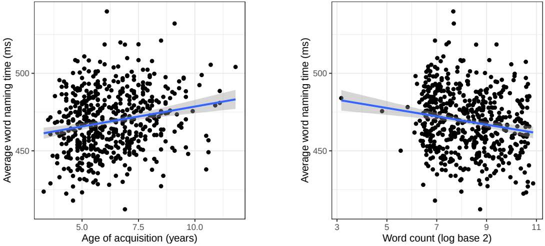
- What predicts word recognition speed: a word's *frequency* or its *age of acquisition*?



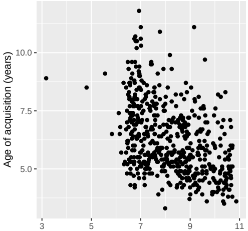
- But look at the relationship between the two!

A problem of credit assignment

- What predicts word recognition speed: a word's *frequency* or its *age of acquisition*?



- But look at the relationship between the two!



A problem of credit assignment

- Linear regression is designed to handle these cases!

$$\mathsf{RT} = \alpha + \beta_{\mathsf{Freq}}\mathsf{Freq} + \beta_{\mathsf{AoA}}\mathsf{AoA} + \epsilon$$

A problem of credit assignment

- Linear regression is designed to handle these cases!

$$\text{RT} = \alpha + \beta_{\text{Freq}}\text{Freq} + \beta_{\text{AoA}}\text{AoA} + \epsilon$$

- So the model matrix will look like:

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

A problem of credit assignment

- Linear regression is designed to handle these cases!

$$\mathsf{RT} = \alpha + \beta_{\mathsf{Freq}}\mathsf{Freq} + \beta_{\mathsf{AoA}}\mathsf{AoA} + \epsilon$$

- So the model matrix will look like:

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

- Remember, we can read off the correlations among the parameter estimates from $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$, which in this case is

$$(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} = \begin{bmatrix} 0.183 & -0.0143 & -0.0102 \\ -0.0143 & 0.00137 & 0.000499 \\ -0.0102 & 0.000499 & 0.000966 \end{bmatrix}$$

A problem of credit assignment

- Linear regression is designed to handle these cases!

$$\text{RT} = \alpha + \beta_{\text{Freq}}\text{Freq} + \beta_{\text{AoA}}\text{AoA} + \epsilon$$

- So the model matrix will look like:

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

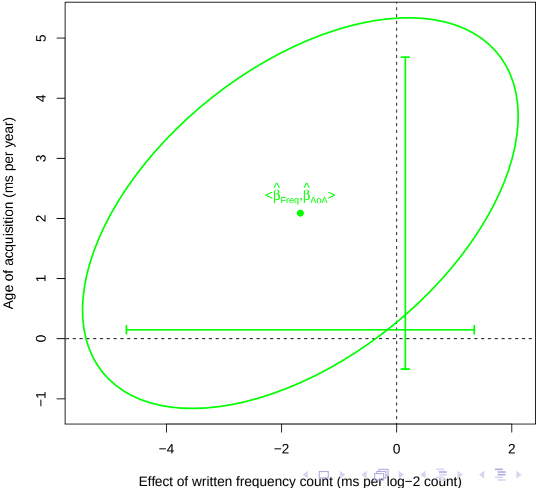
- Remember, we can read off the correlations among the parameter estimates from $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, which in this case is

$$(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \begin{bmatrix} 0.183 & -0.0143 & -0.0102 \\ -0.0143 & 0.00137 & 0.000499 \\ -0.0102 & 0.000499 & 0.000966 \end{bmatrix}$$

- Thus the correlation between $\widehat{\beta}_{\text{Freq}}$ and $\widehat{\beta}_{\text{AoA}}$ is $\frac{0.000499}{\sqrt{0.00137 \times 0.000966}} = 0.4341883$

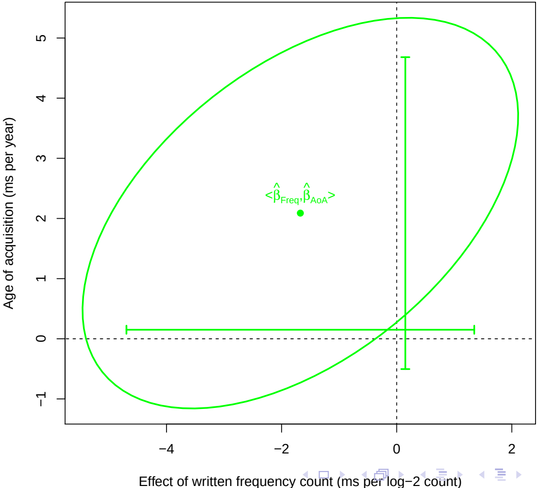
A problem of credit assignment

A 95% confidence region for a random subset of 100 words from the English Lexicon Project (Spieler and Balota, 1997):



A problem of credit assignment

A 95% confidence region for a random subset of 100 words from the English Lexicon Project (Spieler and Balota, 1997):



A problem of credit assignment

We see this in R output as well:

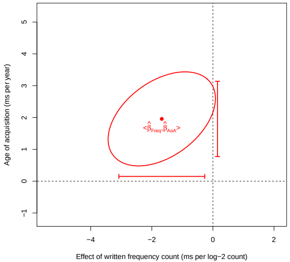
```
summary(m)

##
## Call:
## lm(formula = exp(RTnaming) ~ WrittenFrequencyLog2 + AoA, data = dat.english.subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.429 -12.719   0.439  11.026  41.035
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           471.958     17.893  26.376   <2e-16 ***
## WrittenFrequencyLog2   -1.674      1.523  -1.100    0.274
## AoA                     2.089      1.306   1.599    0.113
## ---
## Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
##
## Residual standard error: 17.12 on 97 degrees of freedom
## Multiple R-squared:  0.07014, Adjusted R-squared:  0.05097
## F-statistic: 3.659 on 2 and 97 DF,  p-value: 0.02939
```

A problem of credit assignment

- Problem of credit assignment: it looks like jointly, *some* combination of frequency and age of acquisition predicts word naming speed

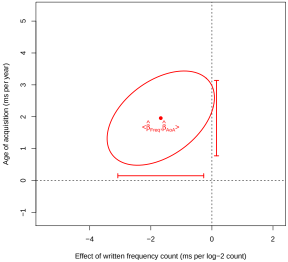
A problem of credit assignment

- Problem of credit assignment: it looks like jointly, *some* combination of frequency and age of acquisition predicts word naming speed
- With more data, of course, we can get a clearer picture—e.g., from complete dataset:



A problem of credit assignment

- Problem of credit assignment: it looks like jointly, *some* combination of frequency and age of acquisition predicts word naming speed
- With more data, of course, we can get a clearer picture—e.g., from complete dataset:



- But suppose we *didn't* have more data: can we test the hypothesis that *some* combination of frequency and age of acquisition predicts naming time?

Nesting of models

- Sometimes, one model will be a more restricted version of another

$$
\begin{aligned}
M_1 &: & y &= \alpha + \beta_{\text{Freq}}\text{Freq} + \epsilon \\
M_2 &: & y &= \alpha + \beta_{\text{Freq}}\text{Freq} + \beta_{\text{AoA}}\text{AoA} + \epsilon
\end{aligned}
$$

Nesting of models

- Sometimes, one model will be a more restricted version of another

$$
\begin{aligned}
M_1: &\quad y = \alpha + \beta_{\text{Freq}}\text{Freq} + \epsilon \\
M_2: &\quad y = \alpha + \beta_{\text{Freq}}\text{Freq} + \beta_{\text{AoA}}\text{AoA} + \epsilon
\end{aligned}
$$

- Here, any version of M_1 can be mimicked by M_2, by setting $\beta_{\text{AoA}} = 0$

Nesting of models

- Sometimes, one model will be a more restricted version of another

$$\begin{aligned} M_1 : \quad & y = \alpha + \beta_{\mathsf{Freq}}\mathsf{Freq} + \epsilon \\ M_2 : \quad & y = \alpha + \beta_{\mathsf{Freq}}\mathsf{Freq} + \beta_{\mathsf{AoA}}\mathsf{AoA} + \epsilon \end{aligned}$$

- Here, any version of M_1 can be mimicked by M_2, by setting $\beta_{\mathsf{AoA}} = 0$
- Under such situations, we say that M_1 is nested inside M_2

Nesting of models

- Sometimes, one model will be a more restricted version of another

$$
\begin{aligned}
M_1 : \quad & y = \alpha + \beta_{\mathsf{Freq}}\mathsf{Freq} + \epsilon \\
M_2 : \quad & y = \alpha + \beta_{\mathsf{Freq}}\mathsf{Freq} + \beta_{\mathsf{AoA}}\mathsf{AoA} + \epsilon
\end{aligned}
$$

- Here, any version of M_1 can be mimicked by M_2, by setting $\beta_{\mathsf{AoA}} = 0$
- Under such situations, we say that M_1 is nested inside M_2
- Sometimes, it will be less apparent that two models are in a nesting relation—but it can be checked automatically!

The decomposition of variance

- Beautiful property of linear models: decomposition of variance. If model M produces fitted values $\{\hat{y}_i\}$, then the overall variance $\mathrm{Var}(y)$ can be decomposed:

$$\begin{aligned}\mathrm{Var}(y) &= \sum_j (y_j - \bar{y})^2 \\ &= \overbrace{\sum_j (y_j - \hat{y}_j)^2}^{\text{unexplained}} + \overbrace{\sum_j (\hat{y}_j - \bar{y})^2}^{\mathrm{Var}_M(y)}\end{aligned}$$

The decomposition of variance

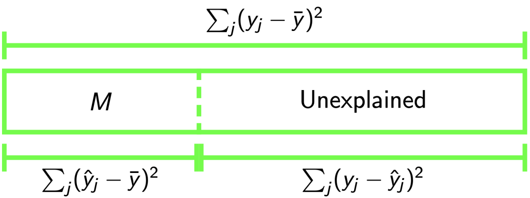
- Beautiful property of linear models: decomposition of variance. If model M produces fitted values $\{\hat{y}_i\}$, then the overall variance $\mathrm{Var}(y)$ can be decomposed:

$$
\begin{aligned}
\mathrm{Var}(y) &= \sum_j (y_j - \bar{y})^2 \\
&= \overbrace{\sum_j (y_j - \hat{y}_j)^2}^{\text{unexplained}} + \overbrace{\sum_j (\hat{y}_j - \bar{y})^2}^{\mathrm{Var}_M(y)}
\end{aligned}
$$

The decomposition of variance

- Beautiful property of linear models: decomposition of variance. If model M produces fitted values $\{\hat{y}_i\}$, then the overall variance $\mathrm{Var}(y)$ can be decomposed:

$$
\begin{aligned}
\mathrm{Var}(y) &= \sum_j (y_j - \bar{y})^2 \\
&= \overbrace{\sum_j (y_j - \hat{y}_j)^2}^{\text{unexplained}} + \overbrace{\sum_j (\hat{y}_j - \bar{y})^2}^{\mathrm{Var}_M(y)}
\end{aligned}
$$



The decomposition of variance

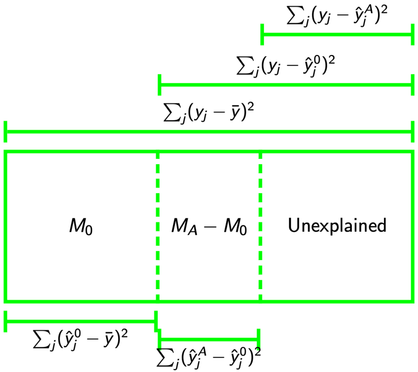
- More generally, take models M_0, M_A, such that M_0 is nested inside M_A

The decomposition of variance

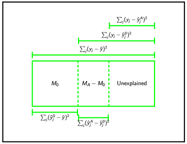
- More generally, take models M_0, M_A, such that M_0 is nested inside M_A
- Then the variance can be decomposed into three pieces:

$$\overbrace{\sum_j (y_j - \bar{y})^2}^{\mathrm{Var}(y)} = \overbrace{\sum_j (\hat{y}_j^0 - \bar{y})^2}^{\mathrm{Var}_0(y)} + \overbrace{\sum_j (\hat{y}_j^A - \hat{y}_j^0)^2}^{\mathrm{Var}_A(y) - \mathrm{Var}_0(y)} + \overbrace{\sum_j (y_j - \hat{y}_j^A)^2}^{\text{unexplained}}$$

The decomposition of variance—graphically represented



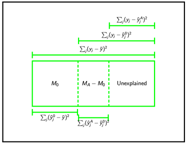
ANOVA: NHST model comparison with the F statistic



$$\begin{array}{ll} \sum_j (y_j - \hat{y}_j^0)^2 & RSS_0 \\ \sum_j (y_j - \hat{y}_j^A)^2 & RSS_A \end{array}$$

- We can compare M_0 and M_A (suppose they respectively have m_0 and m_A predictors) with null-hypothesis significance testing

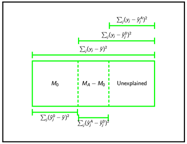
ANOVA: NHST model comparison with the F statistic



$$\begin{array}{ll} \sum_j (y_j - \hat{y}_j^0)^2 & RSS_0 \\ \sum_j (y_j - \hat{y}_j^A)^2 & RSS_A \end{array}$$

- We can compare M_0 and M_A (suppose they respectively have m_0 and m_A predictors) with null-hypothesis significance testing
- Suppose that M_0 is true: then the middle box ($M_A - M_0$) and the right box (Unexplained) are both chi-square distributed, with $m_A - m_0$ and $n - m_A - 1$ degrees of freedom respectively

ANOVA: NHST model comparison with the F statistic



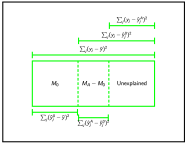
$\sum_j (y_j - \hat{y}_j^0)^2 \quad RSS_0$

$\sum_j (y_j - \hat{y}_j^A)^2 \quad RSS_A$

- We can compare M_0 and M_A (suppose they respectively have m_0 and m_A predictors) with null-hypothesis significance testing
- Suppose that M_0 is true: then the middle box ($M_A - M_0$) and the right box (Unexplained) are both chi-square distributed, with $m_A - m_0$ and $n - m_A - 1$ degrees of freedom respectively
- Thus we have

$$\frac{(RSS_0 - RSS_A)/(m_A - m_0)}{RSS_A/(n - m_A - 1)} \sim F_{m_A - m_0, n - m_A - 1}$$

ANOVA: NHST model comparison with the F statistic



$$\begin{array}{ll} \sum_j (y_j - \hat{y}_j^0)^2 & RSS_0 \\ \sum_j (y_j - \hat{y}_j^A)^2 & RSS_A \end{array}$$

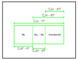
- We can compare M_0 and M_A (suppose they respectively have m_0 and m_A predictors) with null-hypothesis significance testing
- Suppose that M_0 is true: then the middle box ($M_A - M_0$) and the right box (Unexplained) are both chi-square distributed, with $m_A - m_0$ and $n - m_A - 1$ degrees of freedom respectively
- Thus we have

$$\frac{(RSS_0 - RSS_A)/(m_A - m_0)}{RSS_A/(n - m_A - 1)} \sim F_{m_A - m_0, n - m_A - 1}$$

- This hypothesis test is the Analysis of Variance (ANOVA)!

Analysis of Variance: an example

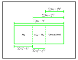
- In our example, formulate the hypothesis: does *some combination of frequency and age of acquisition* meaningfully predict naming time?



$\sum_j (y_j - \hat{y}_j^0)^2$ RSS_0
$\sum_j (y_j - \hat{y}_j^A)^2$ RSS_A

Analysis of Variance: an example

- In our example, formulate the hypothesis: does *some combination of frequency and age of acquisition* meaningfully predict naming time?

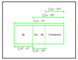


$\sum_j (y_j - \hat{y}_j^0)^2$ RSS_0
$\sum_j (y_j - \hat{y}_j^A)^2$ RSS_A

Analysis of Variance: an example

- In our example, formulate the hypothesis: does *some combination of frequency and age of acquisition* meaningfully predict naming time?

$$
\begin{aligned}
M_0(m_0 = 0) &\quad RT \sim \alpha + \epsilon \\
M_A(m_A = 2) &\quad RT \sim \alpha + \beta_{\text{Freq}}\text{Freq} + \beta_{\text{AoA}}\text{AoA} + \epsilon
\end{aligned}
$$



$\sum_j (y_j - \hat{y}_j^0)^2$ RSS_0
$\sum_j (y_j - \hat{y}_j^A)^2$ RSS_A

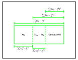
- For us (note that $n = 100$):

$$
\frac{(RSS_0 - RSS_A)/(m_A - m_0)}{RSS_A/(n - m_A - 1)} = \frac{(30589 - 28443)/2}{28443/97}
$$
$$
= 3.66
$$

Analysis of Variance: an example

- In our example, formulate the hypothesis: does *some combination of frequency and age of acquisition* meaningfully predict naming time?

$$
\begin{array}{ll}
M_0(m_0 = 0) & RT \sim \alpha + \epsilon \\
M_A(m_A = 2) & RT \sim \alpha + \beta_{\text{Freq}}\text{Freq} + \beta_{\text{AoA}}\text{AoA} + \epsilon
\end{array}
$$



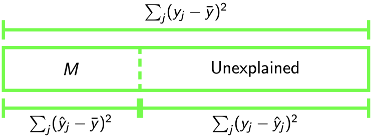
$\sum_j (y_j - \hat{y}_j^0)^2$ RSS_0
$\sum_j (y_j - \hat{y}_j^A)^2$ RSS_A

- For us (note that $n = 100$):

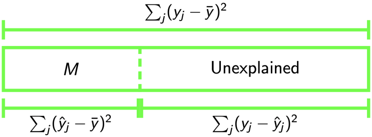
$$
\frac{(RSS_0 - RSS_A)/(m_A - m_0)}{RSS_A/(n - m_A - 1)} = \frac{(30589 - 28443)/2}{28443/97}
$$
$$
= 3.66
$$

- Consulting the cumulative distribution function for $F_{2,97}$, we get a p-value of 0.0294$\rightarrow$yes, there is good evidence that some combination of frequency and AoA predicts RTs!

How much of the variance does your model explain?



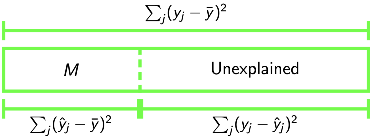
How much of the variance does your model explain?



- The coefficient of determination, R^2, is often used to quantify overall model fit in a scale-independent way

$$R^2 = 1 - \frac{\text{Unexplained variance}}{\text{Total variance}}$$
$$= 1 - \frac{\sum_j (y_j - \hat{y}_j)^2}{\sum_j (y_j - \bar{y})^2}$$

How much of the variance does your model explain?



- The coefficient of determination, R^2, is often used to quantify overall model fit in a scale-independent way

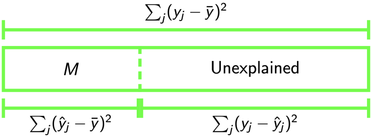
$$R^2 = 1 - \frac{\text{Unexplained variance}}{\text{Total variance}}$$
$$= 1 - \frac{\sum_j (y_j - \hat{y}_j)^2}{\sum_j (y_j - \bar{y})^2}$$

- For the full set of $n = 539$ English words, our model

$$RT \sim \text{Freq} + \text{AoA}$$

has $R^2 = 0.0495$

How much of the variance does your model explain?



- The coefficient of determination, R^2, is often used to quantify overall model fit in a scale-independent way

$$
\begin{aligned}
R^2 &= 1 - \frac{\text{Unexplained variance}}{\text{Total variance}} \\
&= 1 - \frac{\sum_j (y_j - \hat{y}_j)^2}{\sum_j (y_j - \bar{y})^2}
\end{aligned}
$$

- For the full set of $n = 539$ English words, our model

$$RT \sim \text{Freq} + \text{AoA}$$

  has $R^2 = 0.0495$
- $\rightarrow$Although both are useful in predicting RTs, only a small fraction of variance across words in average RT is explained!

References I

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., and Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, 63:489–505.

Spieler, D. H. and Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, 6:411–416.

A note on p-values and philosophy of science

- Frequentist hypothesis testing means the Neyman-Pearson paradigm, with an asymmetry between null (H_0) and alternative (H_1) hypotheses

A note on p-values and philosophy of science

- Frequentist hypothesis testing means the Neyman-Pearson paradigm, with an asymmetry between null (H_0) and alternative (H_1) hypotheses
- A p-value from a dataset D is how unlikely a given dataset was to be produced under H_0

A note on p-values and philosophy of science

- Frequentist hypothesis testing means the Neyman-Pearson paradigm, with an asymmetry between null (H_0) and alternative (H_1) hypotheses
- A p-value from a dataset D is how unlikely a given dataset was to be produced under H_0
- Note that so-called "p_{MCMC}" is NOT a p-value in the Neyman-Pearson sense!

A note on p-values and philosophy of science

- Frequentist hypothesis testing means the Neyman-Pearson paradigm, with an asymmetry between null (H_0) and alternative (H_1) hypotheses
- A p-value from a dataset D is how unlikely a given dataset was to be produced under H_0
- Note that so-called "p_{MCMC}" is NOT a p-value in the Neyman-Pearson sense!
- Weakness, both in practice and in principle: the alternative hypothesis is never actually used (except indirectly in determining optimal acceptance and rejection regions)

A note on p-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)}\frac{P(H_0)}{P(H_1)}$$

A note on p-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)}\frac{P(H_0)}{P(H_1)}$$

- I am fundamentally Bayesian in my philosophy of science

A note on p-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)}\frac{P(H_0)}{P(H_1)}$$

- I am fundamentally Bayesian in my philosophy of science
- But, weakness in practice: your likelihoods $P(D|H_0)$ and $P(D|H_1)$ can depend on fine details of your assumptions about H_0 and H_1

A note on p-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)} \frac{P(H_0)}{P(H_1)}$$

- I am fundamentally Bayesian in my philosophy of science
- But, weakness in practice: your likelihoods $P(D|H_0)$ and $P(D|H_1)$ can depend on fine details of your assumptions about H_0 and H_1
- I do not trust you to assess these likelihoods neutrally! (Nor should you trust me)

A note on p-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)} \frac{P(H_0)}{P(H_1)}$$

- I am fundamentally Bayesian in my philosophy of science
- But, weakness in practice: your likelihoods $P(D|H_0)$ and $P(D|H_1)$ can depend on fine details of your assumptions about H_0 and H_1
- I do not trust you to assess these likelihoods neutrally! (Nor should you trust me)
- So for me, the p-value of your experiment serves as a rough indicator of how small $P(D|H_0)$ may be

A note on p-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)}\frac{P(H_0)}{P(H_1)}$$

- I am fundamentally Bayesian in my philosophy of science
- But, weakness in practice: your likelihoods $P(D|H_0)$ and $P(D|H_1)$ can depend on fine details of your assumptions about H_0 and H_1
- I do not trust you to assess these likelihoods neutrally! (Nor should you trust me)
- So for me, the p-value of your experiment serves as a rough indicator of how small $P(D|H_0)$ may be
- Technically, such a measure doesn't need to be a true Neyman-Pearson p-value (p_{MCMC} falls into this category)