Model Selection: Information Criteria and Sparsity Approaches

CS 6190: Probabilistic Modeling

March 1, 2018

Linear Model Selection

Linear Model:

$$
\begin{aligned}
y &= X\beta + \epsilon \\
&= \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_K\beta_K + \epsilon
\end{aligned}
$$

$$\epsilon \sim N(0, \sigma^2)$$

Linear Model Selection

Linear Model:

$$
\begin{aligned}
y &= X\beta + \epsilon \\
&= \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_K\beta_K + \epsilon
\end{aligned}
$$

$$\epsilon \sim N(0, \sigma^2)$$

Model Selection Problem:

Which regressors, x_i, should we include in the model?

OASIS Brain Data

http://www.oasis-brains.org

```
> head(cdat[,-1])


  M.F Age MMSE CDR eTIV  nWBV RightHippoVol LeftHippoVol
1   F  73   27 0.5 1454 0.708          2896         2801
2   M  74   30 0.0 1636 0.689          2832         2578
3   F  81   30 0.0 1664 0.679          3557         3495
4   M  76   28 0.5 1738 0.719          3052         2770
5   M  82   27 0.5 1477 0.739          3421         3119
6   F  89   30 0.0 1536 0.715          3760         3167
```

MMSE: Mini-Mental State Exam

CDR : Clinical Dementia Rating

eTIV: Estimated Total Intracranial Volume

nWBV: Normalized Whole Brain Volume

OASIS Brain Data

http://www.oasis-brains.org

```
> head(cdat[,-1])

  M.F Age MMSE CDR eTIV  nWBV RightHippoVol LeftHippoVol
1   F  73   27 0.5 1454 0.708          2896         2801
2   M  74   30 0.0 1636 0.689          2832         2578
3   F  81   30 0.0 1664 0.679          3557         3495
4   M  76   28 0.5 1738 0.719          3052         2770
5   M  82   27 0.5 1477 0.739          3421         3119
6   F  89   30 0.0 1536 0.715          3760         3167
```

Hypotheses of interest:

- Hippocampal volume decreases with age

OASIS Brain Data

http://www.oasis-brains.org

```
> head(cdat[,-1])


  M.F Age MMSE CDR eTIV  nWBV RightHippoVol LeftHippoVol
1   F  73   27 0.5 1454 0.708          2896         2801
2   M  74   30 0.0 1636 0.689          2832         2578
3   F  81   30 0.0 1664 0.679          3557         3495
4   M  76   28 0.5 1738 0.719          3052         2770
5   M  82   27 0.5 1477 0.739          3421         3119
6   F  89   30 0.0 1536 0.715          3760         3167
```

Hypotheses of interest:

- Hippocampal volume decreases with age
- Lower hippocampal volume is also associate with cognitive decline (as measured by MMSE, CDR)

OASIS Brain Data

http://www.oasis-brains.org

```
> head(cdat[,-1])


  M.F Age MMSE CDR eTIV  nWBV RightHippoVol LeftHippoVol
1   F  73   27 0.5 1454 0.708          2896         2801
2   M  74   30 0.0 1636 0.689          2832         2578
3   F  81   30 0.0 1664 0.679          3557         3495
4   M  76   28 0.5 1738 0.719          3052         2770
5   M  82   27 0.5 1477 0.739          3421         3119
6   F  89   30 0.0 1536 0.715          3760         3167
```

What models do we use to test these hypotheses?

- Should we include all variables simultaneously (Age, MMSE, CDR)?
- Which covariates should we include (M.F, eTIV, nWBV)?

All models are wrong, but some are useful.

— George Box

Why not include all the variables we have?

Why not include all the variables we have?

1. Danger of overfitting
2. Each parameter we estimate requires more data

Why not just include covariates that have a "significant" effect in the linear model?

Why not just include covariates that have a “significant” effect in the linear model?

Let’s see!

Age Effects Only

```
> g1 = lm(RightHippoVol ~ Age, data = cdat)
> coef(summary(g1))

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  5660.19    361.858  15.642 9.477e-36
Age           -28.85      4.721  -6.111 5.668e-09
```

Age Effects Only

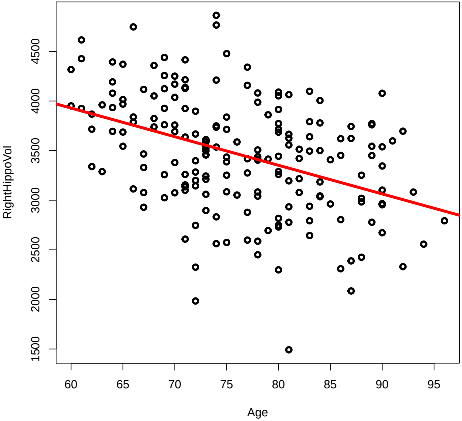
```
> g1 = lm(RightHippoVol ~ Age, data = cdat)
> coef(summary(g1))

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  5660.19    361.858  15.642 9.477e-36
Age           -28.85      4.721  -6.111 5.668e-09
```

Age effect is significant

Age Effects Only

```
> plot(RightHippoVol ~ Age, data = cdat, lwd = 3)
> abline(g1, col = 'red', lwd = 4)
```



Adding Sex Covariate

```
> g2 = lm(RightHippoVol ~ Age + M.F, data = cdat)
> coef(summary(g2))

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  5595.85     362.07  15.455 3.847e-35
Age           -28.62       4.70  -6.088 6.437e-09
M.FM          137.93      81.56   1.691 9.249e-02
```

Adding Sex Covariate

```
> g2 = lm(RightHippoVol ~ Age + M.F, data = cdat)
> coef(summary(g2))

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  5595.85     362.07  15.455 3.847e-35
Age           -28.62       4.70  -6.088 6.437e-09
M.FM          137.93      81.56   1.691 9.249e-02
```

Age effect is significant

Sex effect is significant

Adding Brain Volume Covariate

```
> g3 = lm(RightHippoVol ~ Age + M.F + nWBV, data = cdat)
> coef(summary(g3))

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  -509.99   1045.230 -0.4879 6.262e-01
Age           -10.23      5.228 -1.9570 5.186e-02
M.FM          220.60     75.666  2.9155 3.993e-03
nWBV         6338.75   1029.398  6.1577 4.524e-09
```

Adding Brain Volume Covariate

```
> g3 = lm(RightHippoVol ~ Age + M.F + nWBV, data = cdat)
> coef(summary(g3))

              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  -509.99   1045.230 -0.4879 6.262e-01
Age           -10.23      5.228 -1.9570 5.186e-02
M.FM          220.60     75.666  2.9155 3.993e-03
nWBV         6338.75   1029.398  6.1577 4.524e-09
```

Age effect is NOT significant

Sex effect is significant

Whole brain volume effect is significant

Adding Clinical Dementia Rating

```
> g4 = lm(RightHippoVol ~ Age + M.F + nWBV + CDR, data = cdat)
> coef(summary(g4))

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1828.92   1077.497   1.697 9.133e-02
Age           -15.05      4.982  -3.021 2.878e-03
M.FM          237.85     70.921   3.354 9.692e-04
nWBV         3877.48   1074.262   3.609 3.960e-04
CDR          -496.90     95.796  -5.187 5.632e-07
```

Adding Clinical Dementia Rating

```
> g4 = lm(RightHippoVol ~ Age + M.F + nWBV + CDR, data = cdat)
> coef(summary(g4))

              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1828.92   1077.497   1.697 9.133e-02
Age           -15.05      4.982  -3.021 2.878e-03
M.FM          237.85     70.921   3.354 9.692e-04
nWBV         3877.48   1074.262   3.609 3.960e-04
CDR          -496.90     95.796  -5.187 5.632e-07
```

Everything is significant!

Summary

- Can't choose models based on p-values!

Summary

- Can't choose models based on p-values!
- Statistical significance can be manipulated by inclusion/exclusion of covariates

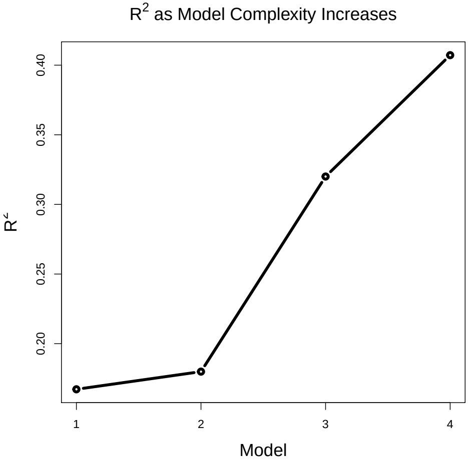
Summary

- Can't choose models based on p-values!
- Statistical significance can be manipulated by inclusion/exclusion of covariates
- Need a systematic and automatic method for selecting models

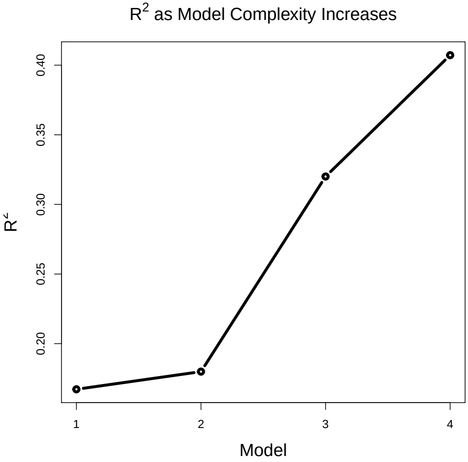
Summary

- Can't choose models based on p-values!
- Statistical significance can be manipulated by inclusion/exclusion of covariates
- Need a systematic and automatic method for selecting models
- Included variables and model selection procedure should be decided before analysis

Highest R^2 or Likelihood?



Highest R^2 or Likelihood?



R^2 always increases when you add covariates

Occam's Razor

Choose the simplest model that explains your data, i.e., the fewest parameters.

Akaike Information Criteria[1]

Pick the model that minimizes

$$\mathrm{AIC} = 2k - 2\ln(L)$$

k: number of parameters

L: log-likelihood

[1] Akaike, IEEE TAC, 1974

Akaike Information Criteria[1]

Pick the model that minimizes

$$\mathrm{AIC} = 2k - 2\ln(L)$$

k: number of parameters
L: log-likelihood

Tradeoff between

maximizing likelihood
and
minimizing number of parameters

[1] Akaike, IEEE TAC, 1974

AIC Under Gaussian Likelihood

If the model has normally-distributed errors,

$$\begin{aligned} \mathrm{AIC} &= 2k - 2\ln(L) \\ &= 2k + n\ln\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i{}^2\right) \end{aligned}$$

$\hat{\epsilon}_i$: estimated residual of ith data point

Motivation of AIC

- We want the best approximation of some "true" density $f(x)$.
- Given candidate models: $g_i(x|\theta_i)$

$$K(f, g_i) = \int f(x) \ln f(x) dx - \int f(x) \ln g_i(x|\theta_i) dx$$

Motivation of AIC

- We want the best approximation of some “true” density $f(x)$.
- Given candidate models: $g_i(x|\theta_i)$
- Minimize the Kullback-Leibler divergence:

$$K(f, g_i) = \int f(x) \ln f(x) dx - \int f(x) \ln g_i(x|\theta_i) dx$$

Motivation of AIC

- We want the best approximation of some “true” density $f(x)$.
- Given candidate models: $g_i(x|\theta_i)$
- Minimize the Kullback-Leibler divergence:

$$K(f, g_i) = \int f(x) \ln f(x) dx - \int f(x) \ln g_i(x|\theta_i) dx$$

- AIC approximates this KL divergence (up to a constant in g_i)

AICc: Bias-corrected AIC

- AIC has a first-order correction for bias

AICc: Bias-corrected AIC

- AIC has a first-order correction for bias
- The bias can still be significant for small n

AICc: Bias-corrected AIC

- AIC has a first-order correction for bias
- The bias can still be significant for small n
- A second-order correction of the bias gives:

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$$

Nice Review Article on AIC

Burnham, K. P.; Anderson, D. R. (2004), “Multimodel inference: understanding AIC and BIC in Model Selection”, Sociological Methods and Research 33: 261-304.

R Package: `FindMinIC`

Install from CRAN:

```
> install.packages("FindMinIC")
```

- Tests all 2^K possible subsets of K regressors
- Ranks them based on AIC (or AICc, or BIC)
- Regressors can be fixed to always be included

OASIS Example Revisited

```
> aicModels = FindMinIC(
+   RightHippoVol ~ Age + CDR + MMSE + M.F + nWBV + eTIV,
+   data = cdat)
> print(summary(aicModels)$table[1:5,])

     AIC  formula
[1,] 2814 "+Age + CDR + eTIV + nWBV"
[2,] 2815 "+Age + CDR + MMSE + eTIV + nWBV"
[3,] 2816 "+Age + CDR + M.F + eTIV + nWBV"
[4,] 2817 "+Age + CDR + M.F + MMSE + eTIV + nWBV"
[5,] 2821 "+CDR + eTIV + nWBV"
```

OASIS Example Revisited

```
> summary(getFirstModel(aicModels))



Call:
lm(formula = RightHippoVol ~ +Age + CDR + eTIV + nWBV, data = tmp.gds)

Residuals:
    Min      1Q  Median      3Q     Max 
-1692.1  -258.1     9.3   285.0  1341.0 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -862.449   1131.062   -0.76   0.4467    
Age          -13.822      4.671   -2.96   0.0035 ** 
CDR         -462.349     89.816   -5.15  6.8e-07 ***
eTIV           1.237      0.201    6.15  4.9e-09 ***
nWBV        5040.797   1033.980    4.88  2.4e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 423 on 183 degrees of freedom
Multiple R-squared:  0.478,Adjusted R-squared:  0.467 
F-statistic:   42 on 4 and 183 DF,  p-value: <2e-16
```

Model Selection via Sparsity

- Idea: force coefficients to zero by penalizing non-zero entries
- Sparse approximation:

  $$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_0.$$

  Using l_0 norm:

  $$\|\beta\|_0 = \text{“number of non-zero elements of } \beta\text{”}$$

- This is an NP-hard optimization problem

The lasso[2]

- The l_1 norm is a convex relaxation of the l_0 norm:

$$||\beta||_1 = \sum_{i=1}^{K} |\beta_i|$$

- The lasso estimator is

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||^2 + \lambda ||\beta||_1$$

- This is now a convex optimization problem

[2] *Tibshirani, J. Royal. Statist. Soc B., 1996*

Adaptive Sparsity[3]

- Hierarchical prior on β:

$$\beta \sim N(0, \tau)$$
$$\tau \propto \frac{1}{\tau}$$

[3] *Figueiredo, PAMI 2003*

Adaptive Sparsity[3]

- Hierarchical prior on β:

$$\beta \sim N(0, \tau)$$
$$\tau \propto \frac{1}{\tau}$$

- Parameter-free Jeffreys' hyperprior on τ

[3] *Figueiredo, PAMI 2003*

Adaptive Sparsity[3]

- Hierarchical prior on β:

$$\beta \sim N(0, \tau)$$
$$\tau \propto \frac{1}{\tau}$$

- Parameter-free Jeffreys' hyperprior on τ
- MAP estimation of β by EM algorithm

[3] *Figueiredo, PAMI 2003*

Adaptive Sparsity[3]

- Hierarchical prior on β:

$$\beta \sim N(0, \tau)$$
$$\tau \propto \frac{1}{\tau}$$

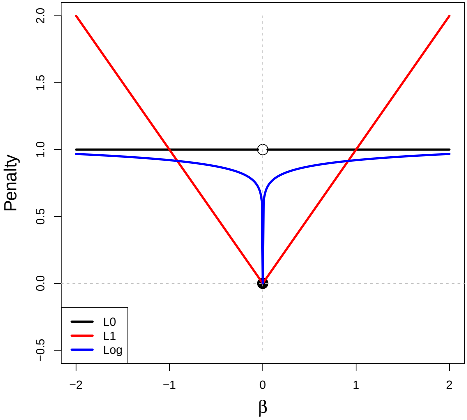
- Parameter-free Jeffreys' hyperprior on τ
- MAP estimation of β by EM algorithm
- After marginalizing τ, equivalent to a log penalty:

$$\log p(\beta) \propto \log(|\beta| + \delta) - \log(\delta)$$

(Need the $\delta > 0$ fudge factor for numerics)

[3] *Figueiredo, PAMI 2003*

Comparison of Penalty Functions



R Package: AdaptiveSparsity

Install from CRAN:

```
> install.packages(AdaptiveSparsity)
```

- Implements Figueiredo's adaptively sparse linear regression (`aslm`)
- Also has a method for estimating sparse Gaussian graphical models (`asggm`)[4]

[4] *Wong, Awate, Fletcher, ICML 2013*

OASIS Example Re-Revisited

```
> g = aslm(
+   RightHippoVol ~ Age + CDR + MMSE + M.F + nWBV + eTIV,
+   data = cdat)
> as.matrix(coef(g))


                  [,1]
(Intercept)      0.000
Age            -15.619
CDR           -477.447
MMSE             0.000
M.FM             0.000
nWBV          4284.793
eTIV             1.126
```

OASIS Example Re-Revisited

```
> g = aslm(
+   RightHippoVol ~ Age + CDR + MMSE + M.F + nWBV + eTIV,
+   data = cdat)
> as.matrix(coef(g))


                  [,1]
(Intercept)      0.000
Age            -15.619
CDR           -477.447
MMSE             0.000
M.FM             0.000
nWBV          4284.793
eTIV             1.126
```

Same coefficients chosen by AIC!

An Interesting Connection

Sparse approximation is **equivalent** to AIC!

An Interesting Connection

Sparse approximation is **equivalent** to AIC!

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_0$$

An Interesting Connection

Sparse approximation is **equivalent** to AIC!

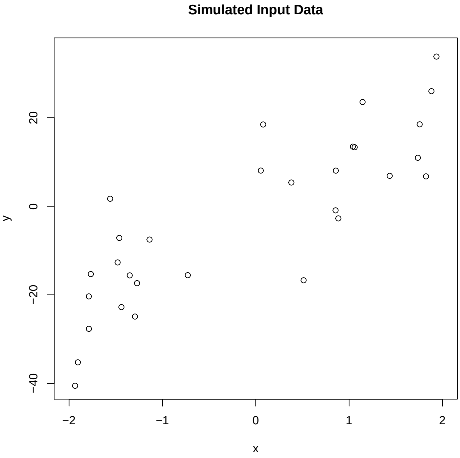
$$
\begin{aligned}
\hat{\beta} &= \arg\min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|_0 \\
&= \arg \min_{k, \|\beta\|_0 = k} -2 \ln L(\beta | y) + 2k, \quad \text{(setting } \lambda = 2\text{)}
\end{aligned}
$$

An Interesting Connection

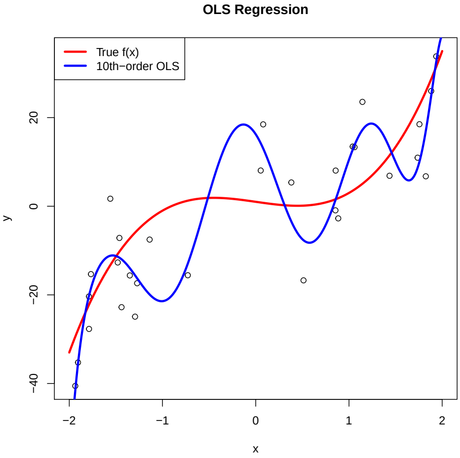
Sparse approximation is **equivalent** to AIC!

$$
\begin{aligned}
\hat{\beta} &= \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_0 \\
&= \arg\min_{k, \|\beta\|_0 = k} -2\ln L(\beta|y) + 2k, \quad \text{(setting } \lambda = 2\text{)} \\
&= \arg\min_{k, \|\beta\|_0 = k} \mathrm{AIC}(\beta)
\end{aligned}
$$

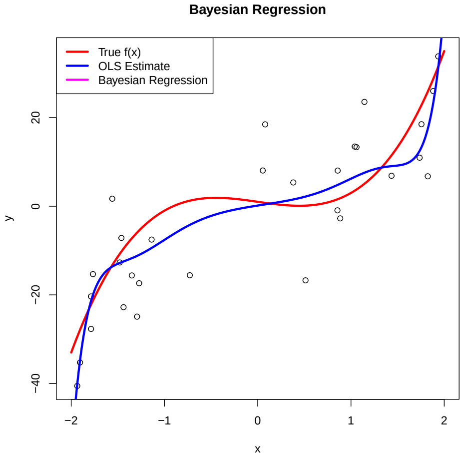
Polynomial Regression Example Revisited



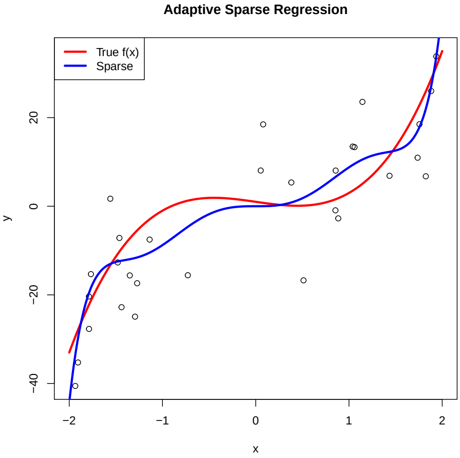
Ordinary Least Squares



Bayesian (Ridge) Regression



Adaptive Sparse Regression



Comparing Coefficient Estimates

```
> print(betaHat)

 [1]   16.238  -32.289 -107.265  106.119  152.427  -76.752
 [7]  -86.301   20.668   21.362   -1.805   -1.936

> print(betaPost)

 [1]  0.156257  3.350884 -0.737490  3.461802 -0.156540
 [6]  1.778690  0.058276 -2.150563 -0.004735  0.458031
[11] -0.002479

> print(betaSparse)

 [1]  0.000  0.000  0.000 16.616  0.000 -9.536  0.000
 [8]  1.692  0.000  0.000  0.000
```