Things Linear
(like algebra and models and stuff)

Schwartz

July 31, 2016

0 Synopsis

Glorified cheat sheet

1 LinAlg

1.1 inner product – *dot product*

$||x||_2, ||x||_1$, and $||x||_\infty$

$$A_{n\times m} \cdot B_{m\times p} \quad = \quad C_{n\times p}$$

$$\begin{bmatrix} \xrightarrow{r} \\ \vdots \\ \longrightarrow \end{bmatrix} \cdot \begin{bmatrix} \downarrow c & \cdots & \downarrow \end{bmatrix} = \begin{bmatrix} C_{r,c} = dot\left(A_r, B_c\right) \end{bmatrix}$$

1.2 outer product – *broadcasing*

$$\begin{bmatrix} \downarrow \\ \downarrow \\ \downarrow \end{bmatrix}_{n\times 1}^{A} \quad \left[\longrightarrow \right]_{1\times m}^{B} \quad = \quad \begin{bmatrix} C_{r,c} = A_r \times B_c \end{bmatrix}_{n\times m}$$

1.3 matrix things and such

- MM is !*communative* but is associative and distributive
- $(AB)^T = B^T A^T$
- If $A = A^T$ then A is *symmetric*

  $\frac{1}{2}A + \frac{1}{2}A^T$ is symmetric
- $\text{tr}ABC = \text{tr}BCA = \text{tr}CAB$
- $\nabla_x a^T x = a$, and if A symmetric

  $\nabla_x x^T Ax = 2Ax$

  $\nabla_x^2 x^T Ax = 2A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$

1.4 Spaces

- $\text{span}\{x_1, \cdots, x_m\} = \mathcal{R}\left(A = [x_1, \cdots, x_m]\right)$

  $$= \left\{ v \in \mathbb{R}^n : v = \sum_{i=1}^{m} \alpha_i x_i \right\}$$

  with $x_i \in \mathbb{R}^n$ and $\alpha_i \in \mathbb{R}$

  Relevant for LM's:

  The *projection* of y onto $\mathcal{R}(A = [x_1, \cdots, x_m])$ is

  $$\min_{v \in \text{span}\{x_1, \cdots, x_m\}} ||y - v||_2 = A(A^T A)^{-1} A^T y$$

  Relevant for SVM's:

  The *projection* of y onto a is $\frac{aa^T}{a^T a} y$
- The *nullspace* of $A \in \mathbb{R}^{n \times m}$ $\mathcal{N}(A) = \{x \in \mathbb{R}^m : Ax = \mathbf{0}\}$

  Why are $\mathcal{R}(A^T)$ and $\mathcal{N}(A)$ *orthogonal complements*?

  I.e., $\{w : w = u + v, u \in \mathcal{R}(A^T), v \in \mathcal{N}(A)\} = \mathbb{R}^m$

  and $\mathcal{R}(A^T) \cap \mathcal{N}(A) = \{\mathbf{0}\}$

1.5 Invertability

- column rank is row rank is rank
- $\text{rank}(A_{(n,m)}) \leq \min(m, n)$, *full* if equal
- $A^{-1}A = I$ if $A \in \mathbb{R}^{n \times n}$ is *full rank*

  otherwise A is *singular/non-invertable*
- $(A^{-1})^T = (A^T)^{-1}$
- $U^T U = I$ $(= UU^T)$ for *orthonormal* U
- $|det\ A| = \text{abs}(|A|)$ is the volume of the $\mathbb{R}^n$-parallelotope formed by the vectors of A
- $|A| = 0$ if A is not full rank

  I.e., A singular (non-invertable)
- $|A^{-1}| = |A|^{-1}$, $|A| = |A^T|$, $|AB| = |A||B|$
- $A = A^T$ is *positive semidefinite* if $x^T Ax \geq 0$

  and *positive definite* if $x^T Ax > 0$

  $\implies$ full rank (i.e., *non-singular/nvertable*)
- For full rank $A_{(n,m)}$ with $n > m$

  the *Gram matrix* $A^T A$ is positive definite

  Relevant for LM's:

  $X^T X$ is inverted in the least squares fit

  $$X(X^T X)^{-1} X^T y$$
- $A^{-1} = \frac{1}{|A|}\text{adj}(A) \in \mathbb{R}^{n \times n}$

  $\text{adj}(A)_{i,j} = (-1)^{i+j} |A_{-i,-j}|$

1.6 Eigens

- An *eigenvector* ($x \neq \mathbf{0}$ in $\mathbb{C}^n$) and *eigenvalue* (λ) pair of $A \in \mathbb{R}^{n \times n}$ satisfy

$$Ax = \lambda x$$

- Solutions for $(\lambda I - A)x = 0$ exist if $(\lambda I - A)$ is singular/the nullspace $\mathcal{N}(\lambda I - A) \neq \{\mathbf{0}\}$

[so that $\text{rank}\big((\lambda I - A)^T\big) = \text{rank}(\lambda I - A)$ is not full]

- In which case eigenvalue solutions to $|\lambda I - A| = 0$ can be used to solve for eigenvectors in

$$(\lambda I - A)x = 0$$

- And for eigenvalues $\lambda_1, \cdots \lambda_n$ of A we have that

$$\text{tr}A = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad |A| = \prod_{i=1}^{n} \lambda_i \quad \text{and} \quad \text{rank}(A) = \sum_{i=1}^{n} 1_{[\lambda_i \neq 0]}$$

- Quiz: what are the eigens for (i) diagonal matrix D and (ii) non-singular A?

- For $A = A^T$ we have that (i) $\lambda_1, \cdots \lambda_n \in \mathbb{R}^n$ and (ii) the eigenvectors are *orthonormal*

  Relevant for MVN:
  All the eigenvalues of Σ must be non zero so that Σ^{-1} exists

- For linearly independent eigenvectors $X = [x_1, \cdots x_n]$ and associated eigenvalues $\Lambda = \text{diag}(\lambda_1, \cdots \lambda_n)$

$$AX = X\Lambda \text{ implies that } A = X\Lambda X^{-1} \text{ is } \textit{diagonalizable}$$

  Relevant for PCA:
  Diagonalization a.k.a. *spectral* or *eigenvalue* decomposition is a special case of the *singular value decomposition* which uses the covariance/correlation matrix rather than the data matrix

- Quiz: A^{-1} for diagonalizable A? What sign are the eigenvalues of positive definite A?

2 Regression

The range of the (full rank) n samples p predictors covariate (or design) matrix $\mathcal{R}(X)$ is the space (of dimension $p < n$) of all possible predictions of $y \in \mathbb{R}^n$

$$\mathcal{Y} = X\boldsymbol{\beta}, \text{ for } \boldsymbol{\beta} \in \mathbb{R}^p$$

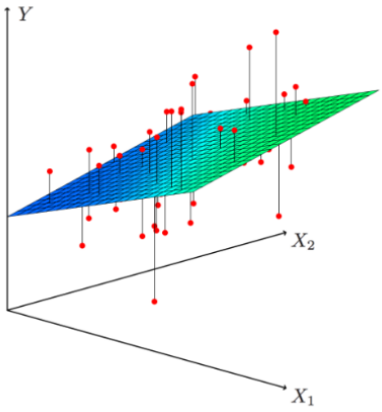
The projection of y onto $\mathcal{R}(X)$ is $\hat{y}$ is

$$\begin{aligned} \min_{X\boldsymbol{\beta}} ||X\boldsymbol{\beta} - y||_2 &= \min_{X\boldsymbol{\beta}} \left((X\boldsymbol{\beta} - y)^T (X\boldsymbol{\beta} - y) \right) \\ &= \min_{X\boldsymbol{\beta}} \left(\boldsymbol{\beta}^T X^T X \boldsymbol{\beta} - 2X^T y \boldsymbol{\beta} + y^T y \right) \end{aligned}$$

which (by taking the derivative) is maximized at

$$\boldsymbol{\beta} = \left(X^T X\right)^{-1} X^T y$$

2.1 Linear Models



$$y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i,\ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$
$$\hat{y}_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Assumptions – when do they matter?
- Higher order terms – what does "linear" mean?
- Challenges interpreting coefficients?
- The role of rank(X) and multicollinearity?
- Significance testing and model building?

2.2 Fit

Total Variation	Total Error	Average Error	Proportion Modeled
$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$RSE = \sqrt{\frac{1}{n-p-1}RSS}$ $= \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p-1}}$	$R^2 = \frac{TSS-RSS}{TSS}$ $= 1 - \frac{RSS}{TSS}$

OLS Regression Results

Dep. Variable:	y	R-squared:	0.933
Model:	OLS	Adj. R-squared:	0.928
Method:	Least Squares	F-statistic:	211.8
Date:	Mon, 03 Nov 2014	Prob (F-statistic):	6.30e-27
Time:	14:45:06	Log-Likelihood:	-34.438
No. Observations:	50	AIC:	76.88
Df Residuals:	46	BIC:	84.52
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	0.4687	0.026	17.751	0.000	0.416	0.522
x2	0.4836	0.104	4.659	0.000	0.275	0.693
x3	-0.0174	0.002	-7.507	0.000	-0.022	-0.013
const	5.2058	0.171	30.405	0.000	4.861	5.550

Omnibus:	0.655	Durbin-Watson:	2.896
Prob(Omnibus):	0.721	Jarque-Bera (JB):	0.360
Skew:	0.207	Prob(JB):	0.835
Kurtosis:	3.026	Cond. No.	221.

Proportion of Variance Explained by model is 93.3%

Measure of the significance of the fit …my model isn't utterly useless ☺

There is an approximately 95% chance that [0.275, 0.693] will contain the true value of β_2

Each coefficient is really significant. Can also think of this as a Partial F-test.

"The average effect on Y of a one unit increase in X_2, holding all other predictors (X_1 & X_3) fixed, is 0.4836"

- However, interpretations are generally pretty hazardous due to correlations among predictors.
- p-values for each coefficient ≈ 0, so might be okay here

Note: Magnitude of the Beta coefficients is NOT how to determine whether predictor contributes. Why?

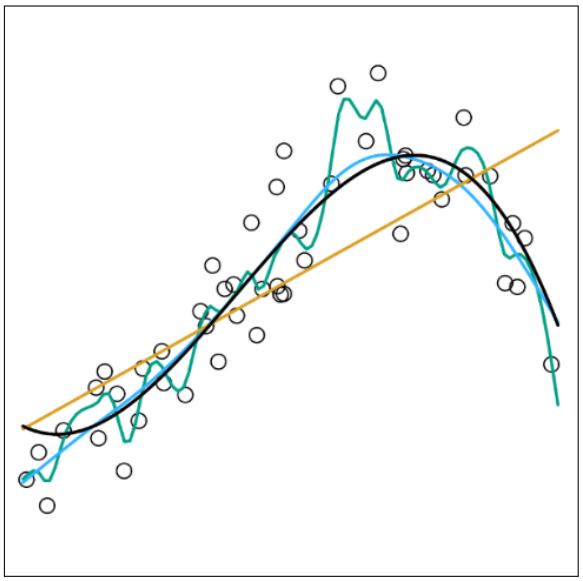
3 Overfitting

Let $y_0 = \theta + \epsilon_0$ with $\theta = f(x_0)$ and $\epsilon \sim \mathrm{N}(0, \sigma_\epsilon^2)$

For estimator $\hat{\theta} = \hat{f}(x_0)$,

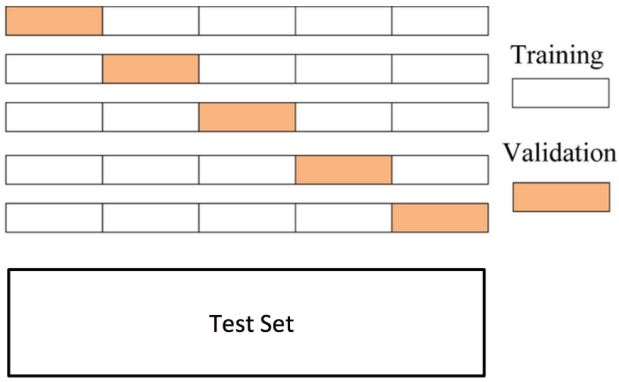
$$\begin{aligned} RSE &= \mathrm{E}\left[\left(y_0 - \hat{\theta}\right)^2\right] \\ &= \sigma_{\hat{\theta}}^2 + \left(\mathrm{E}[\hat{\theta}] - \theta\right)^2 + \sigma_\epsilon^2 \\ &= \text{Variance} + \text{Bias}^2 + \text{Noise} \end{aligned}$$

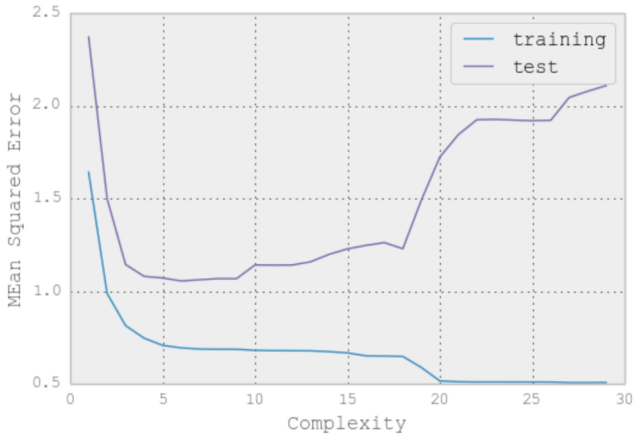
$$\begin{aligned} MSE &= \mathrm{E}\left[\left(\hat{\theta} - \theta\right)^2\right] \\ &= \sigma_{\hat{\theta}}^2 + \left(\mathrm{E}[\hat{\theta}] - \theta\right)^2 \\ &= \text{Variance} + \text{Bias}^2 \end{aligned}$$



The variance/bias tradeoff

3.1 K-folds cross-validation





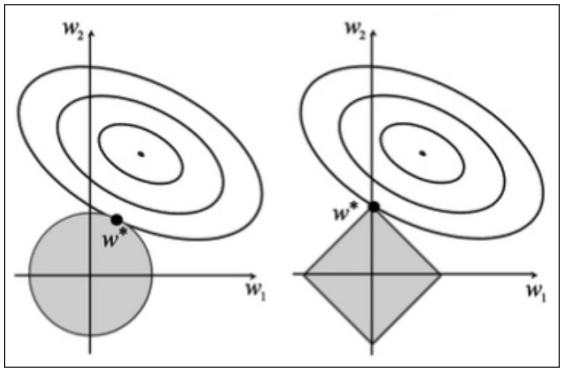
3.2 Regularization

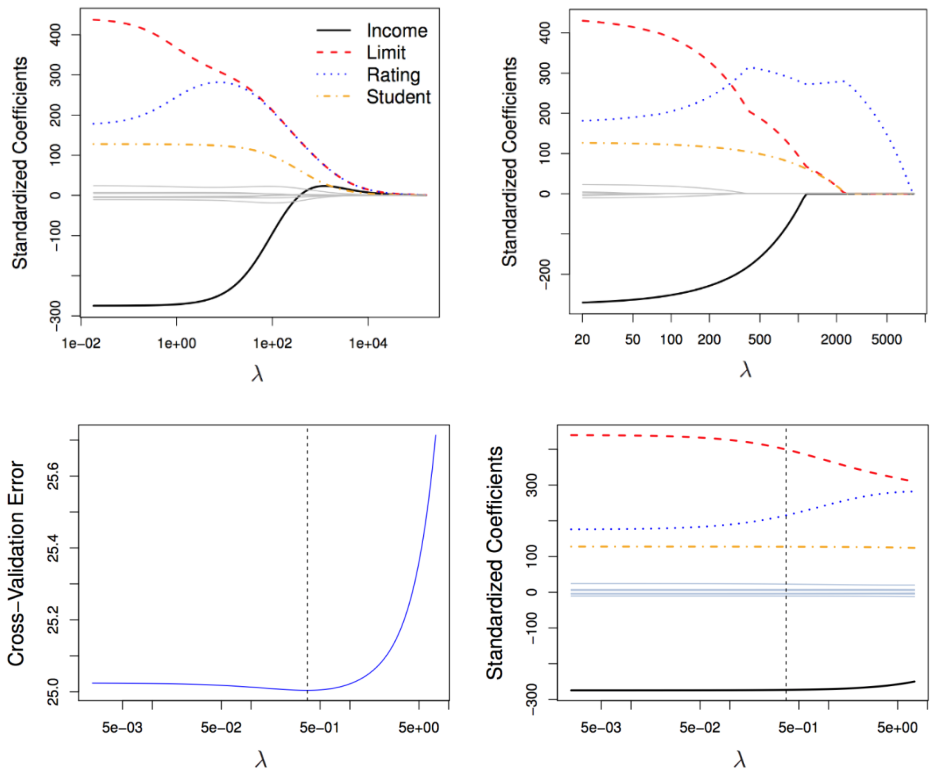
$$\text{LS} = \min_{\boldsymbol{\beta}} ||X\boldsymbol{\beta} - y||_2 \qquad \text{(no penalty)}$$

$$\text{Ridge} = \min_{\boldsymbol{\beta}} ||X\boldsymbol{\beta} - y||_2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (L_2\text{penalty})$$

$$\text{Lasso} = \min_{\boldsymbol{\beta}} ||X\boldsymbol{\beta} - y||_2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (L_1\text{penalty})$$

Penalized coefficients *not* scale equivariant!



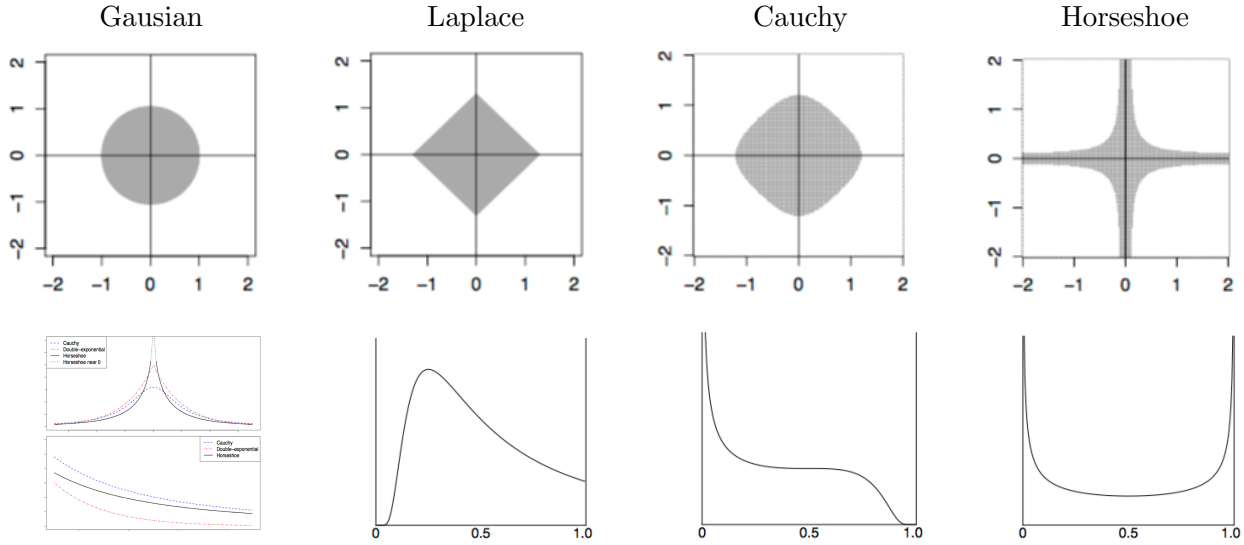


Ridge (top left), Lasso (top right), and cross-validation tuning parameter selection (bottom row)

3.3 Summary

1. Model selection 2. Regularization 3. Dimension reduction (e.g., *principal components reduction*)

3.4 Bonus: Bayesian regularization priors



Tails Implied shrinkage prior profiles from none (0) to total (1) shrinkage

4 Logistic Regression

- The logit *link function* $g(p) = \log\left(\frac{p}{1-p}\right)$ maps $p \in [0,1] \mapsto Z \in \mathbb{R}$
- For a binary outcome Y, setting $\mathrm{E}[Y] = g^{-1}(Z) = \frac{\exp(Z)}{1+\exp(Z)}$ and $Z = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$

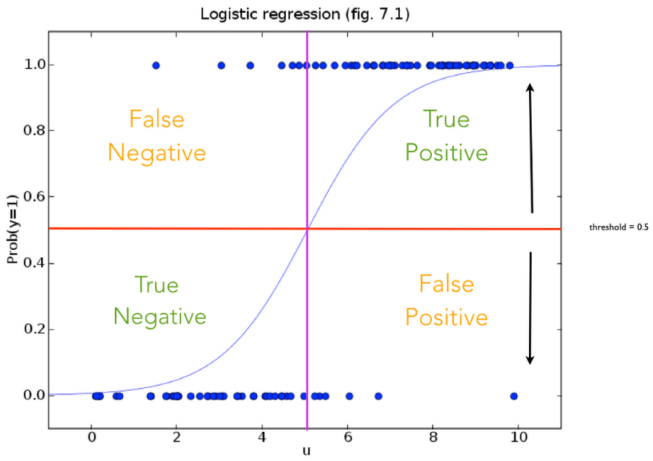
$$Pr(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_n X_m)}}$$

and

$$\exp(\beta_0)\exp(\beta_1 X_1)\cdots\exp(\beta_m X_m) = \frac{Pr(Y = 1|X)}{Pr(Y = 0|X)}$$

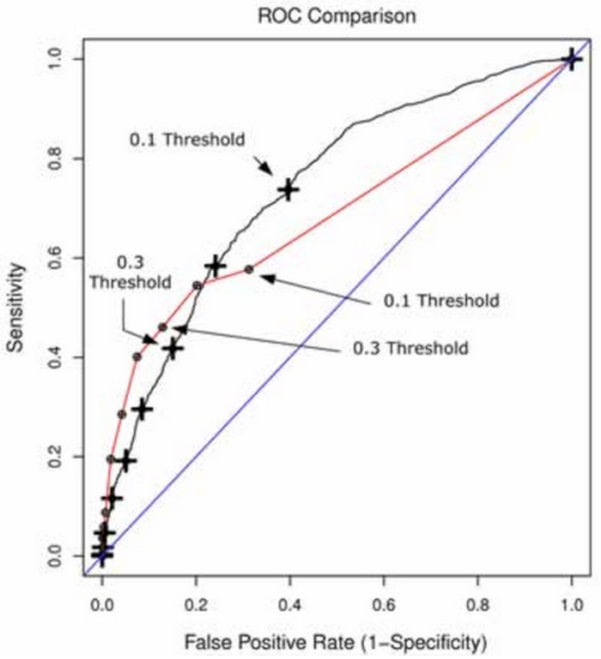
with $\exp(\beta_j)$ the multiplicative (logarithmic scale) *odds* increase for a 1-unit increase in X_j

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
  (Type I error)
- False Negative (FN)
  (Type II error, 1 - power)



- Accuracy: $(TP + TN)/(TP + FP + FN + TN)$
- F1 score: $2TP/(2TP + FP + FN)$

- Sensitivity: $TP/(TP + FN)$
  (power)
- Specificity: $TN/(TN + FP)$
  (1 - Type I error rate α)
- Positive Predictive Value: $TP/(TP + FP)$
  (Precision)
- Negative Predictive Value: $TN/(TN + FN)$



- False Positive Rate: $FP/(FP + TN)$
  (fall-out)
- False *Discovery* Rate: $FP/(FP + TP)$
  (1 - precision)
- False *Negative* Rate: $FN/(FN + TP)$
  (1 - sensitivity)

Appendix: classical model selection

$$C_p = \frac{1}{n}(RSS + 2p\hat{\sigma}^2)$$

Mallow's C_p
p is the total # of parameters
$\hat{\sigma}^2$ is an estimate of the variance of the error, ε

$$AIC = -2logL + 2 \cdot p$$

L is the maximized value of the likelihood function for the model estimated

$$BIC = \frac{1}{n}(RSS + log(n)p\hat{\sigma}^2)$$

This is AIC, except 2 is replaced by log(n). log(n) > 2 for n>7, so BIC generally exacts a heavier penalty for more variables

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

Similar to R^2, but pays price for more variables

Can show AIC and Mallow's C_p are equivalent for linear case