Midterm: Wed 10/18 in class

- **Question answering session Tues 10/17 5:00 pm**
  - Annenberg 110
- **The midterm will allow one two-sided "cheat sheet"**
  - Otherwise closed book, closed notes, no laptop or phone.

Homework

- **For team homework, please only submit *one* copy**
  - If you resubmit, it should be from the same person
  - Otherwise we have two submissions from you

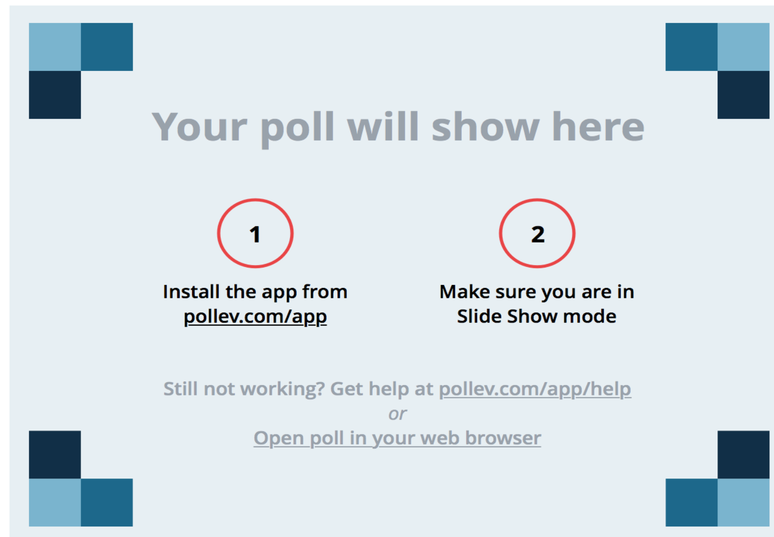
Regression penalty methods

Ridge regression ("Tikhonov regularization") minimizes

$Err + \lambda \, |w|_2^{\,2}$

Is *Err* **here**

- **A)** $\Sigma_i \, (y_i - \hat{y}_i)^2$
- **B)** $(1/n) \, \Sigma_i \, (y_i - \hat{y}_i)^2$
- **C)** $\mathrm{sqrt}(\, (1/n) \, \Sigma_i \, (y_i - \hat{y}_i)^2 \,)$
- **D)** $\mathrm{sqrt}(\Sigma_i \, (y_i - \hat{y}_i)^2 \,)$



Regression penalty methods

Elastic net **regularization minimizes**

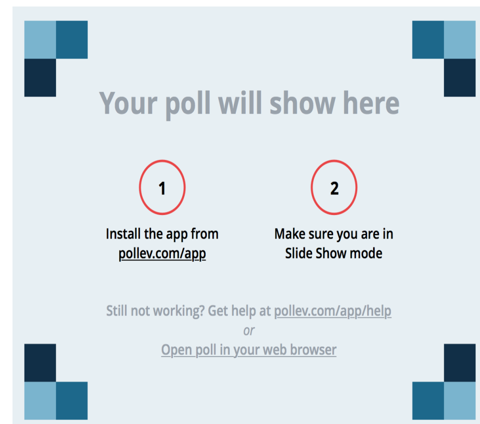
$$Err + \lambda_1 |w|_1 + \lambda_2 |w|_2^{\,2}$$

Will this sometimes zero out some features?

A) yes

B) no

When might this be better than pure L_1?



Regression penalty methods

AIC, BIC and RIC Minimize

$\text{Err}/\, 2\sigma^2 + \lambda\, |w|_0$

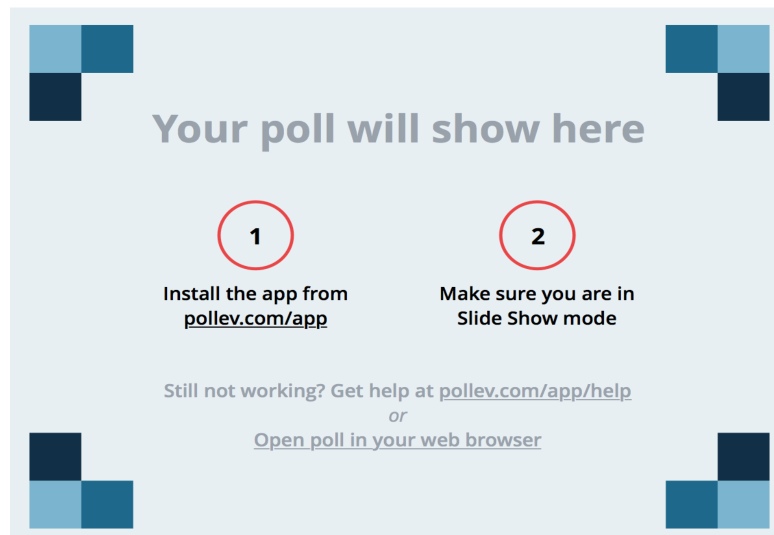
When we don't know σ^2, $\text{Err}/\, 2\sigma^2$ is proportional to

A) $\log(\Sigma_i\, (y_i-\hat{y}_i)^2)$

B) $n \log(\Sigma_i\, (y_i-\hat{y}_i)^2)$

C) $n \log((1/n)\, \Sigma_i\, (y_i-\hat{y}_i)^2)$

D) none of the above



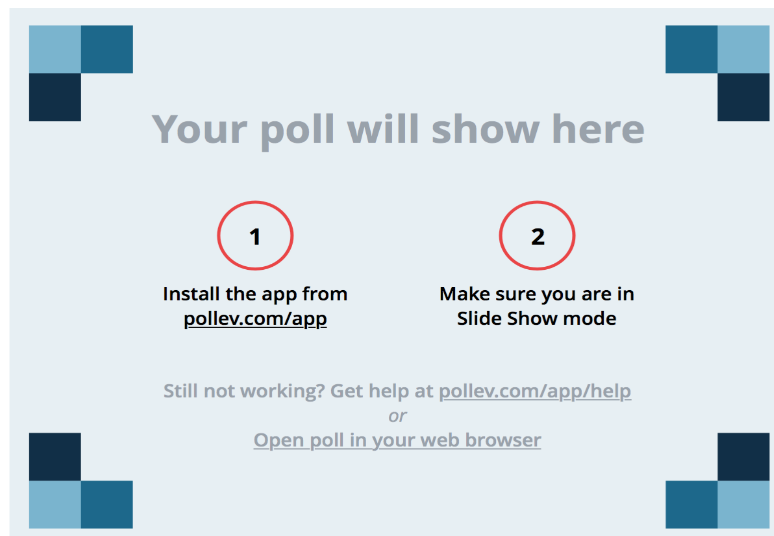
Regression penalty methods

AIC, BIC and RIC Minimize

$$Err / 2\sigma^2 + \lambda \, |w|_0$$

As n becomes large, there is

- **A)** more shrinkage
- **B)** less shrinkage
- **C)** no change



Entropy review

- You need to transmit a sequence of *n* binary observations (e.g. *y* values), which will be
  - "1" with probability $p_1 = 1/8$
  - "0" with probability $p_0 = 7/8$
- What is the minimum number of bits to code the sequence (for large n)?

Entropy review

- **You are doing feature selection where there are far more possible features than observations and expect that roughly 1/8 of the *p* features should be selected.**
- **What would be a better alternative to RIC?**

  $Err/2\sigma^2 + q \log (p)$

How would you code a decision tree

- Assume p = *16* binary variables x
- Binary y n=64 $|y|_0 = 32$ $|y-\hat{y}|_0 = 2$

$$
\begin{array}{ccccc}
 & & x_1 & & \\
 & T/ & & \backslash F & \\
y=1 & & & & x_{14} \\
 & & & T/ & & \backslash F \\
 & & & y=0 & & y=1
\end{array}
$$

How many bits to code the residual?
How many bits to code the decision tree?

Regression penalty methods

Which estimator is *consistent*?

A) AIC

B) BIC

C) RIC

D) none of them

Cross Validation

- **Does LOOCV systematically _____ test error**
  - **A)** Overestimate
  - **B)** Underestimate
  - **C)** Neither

- **Why use Train, Validation and Test sets?**