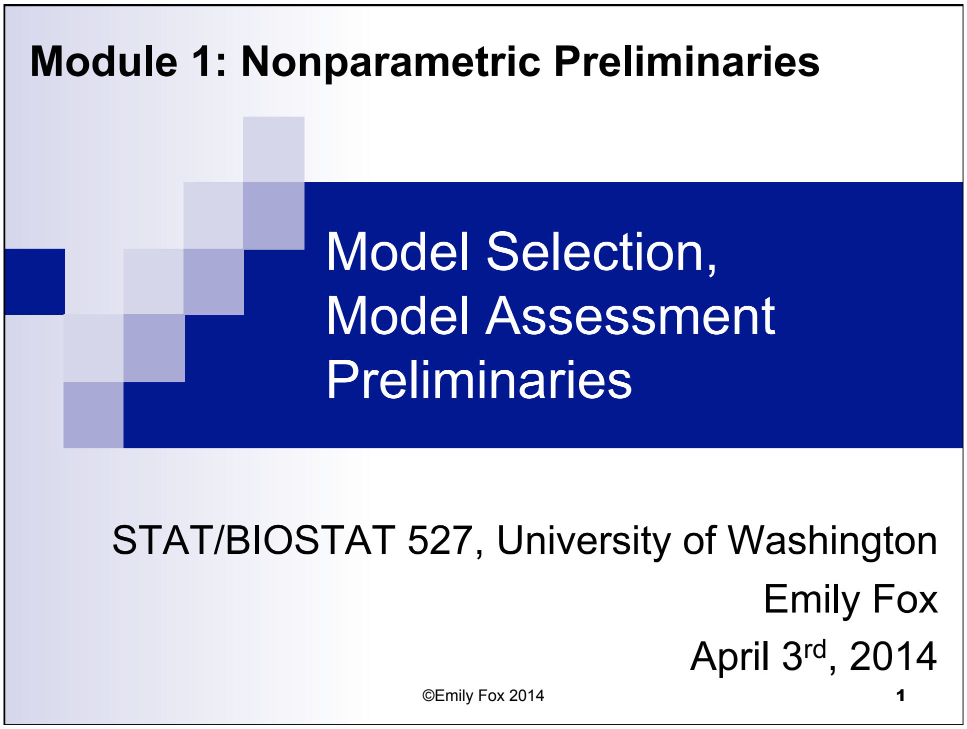


Module 1: Nonparametric Preliminaries

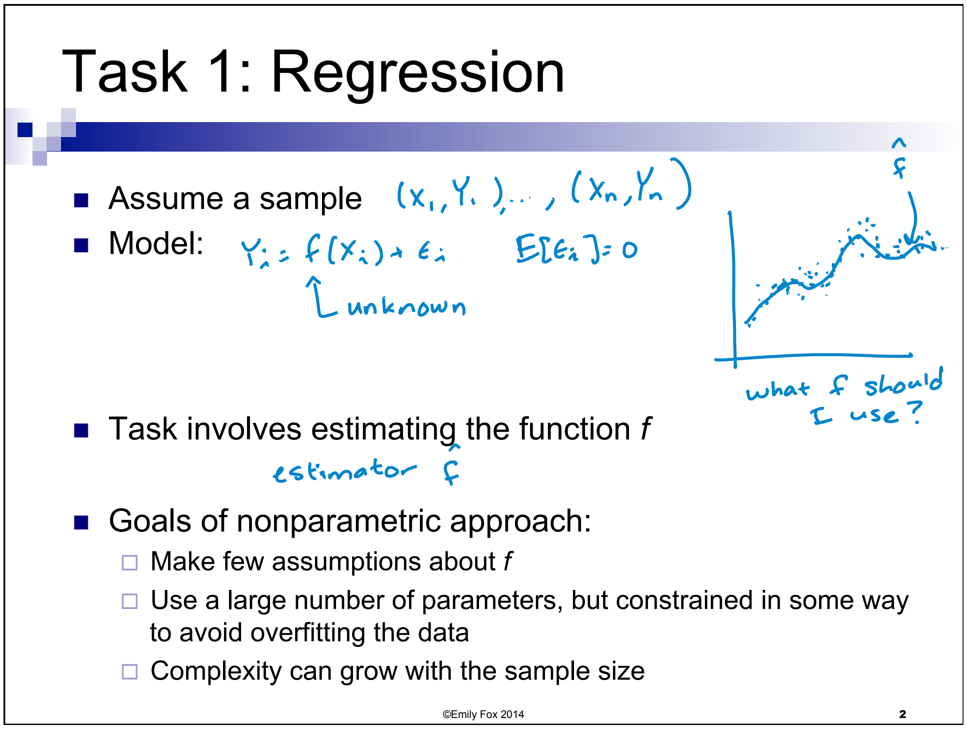
Model Selection, Model Assessment Preliminaries

STAT/BIOSTAT 527, University of Washington

Emily Fox

April 3rd, 2014

©Emily Fox 2014 1



Task 1: Regression

- Assume a sample $(x_1, Y_1), \ldots, (x_n, Y_n)$
- Model: $Y_i = \underset{\uparrow \text{ unknown}}{f}(x_i) + \epsilon_i \qquad E[\epsilon_i] = 0$

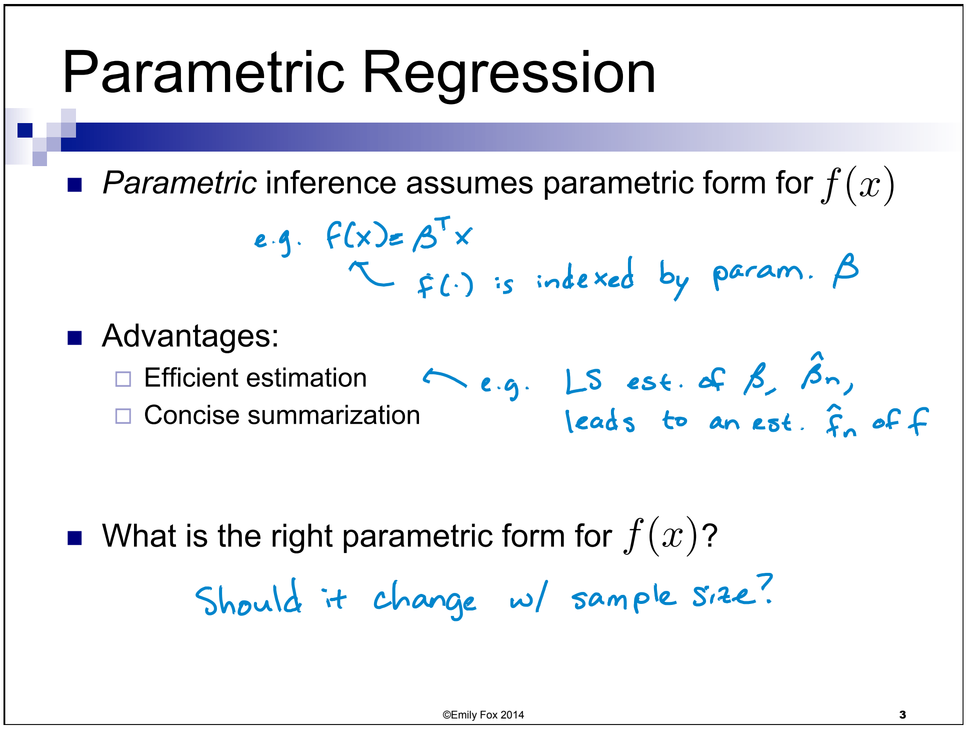
what f should I use?

- Task involves estimating the function *f*

  estimator $\hat{f}$

- Goals of nonparametric approach:
  - □ Make few assumptions about *f*
  - □ Use a large number of parameters, but constrained in some way to avoid overfitting the data
  - □ Complexity can grow with the sample size

©Emily Fox 2014 2



Parametric Regression

- *Parametric* inference assumes parametric form for $f(x)$

  e.g. $f(x) = \beta^T x$

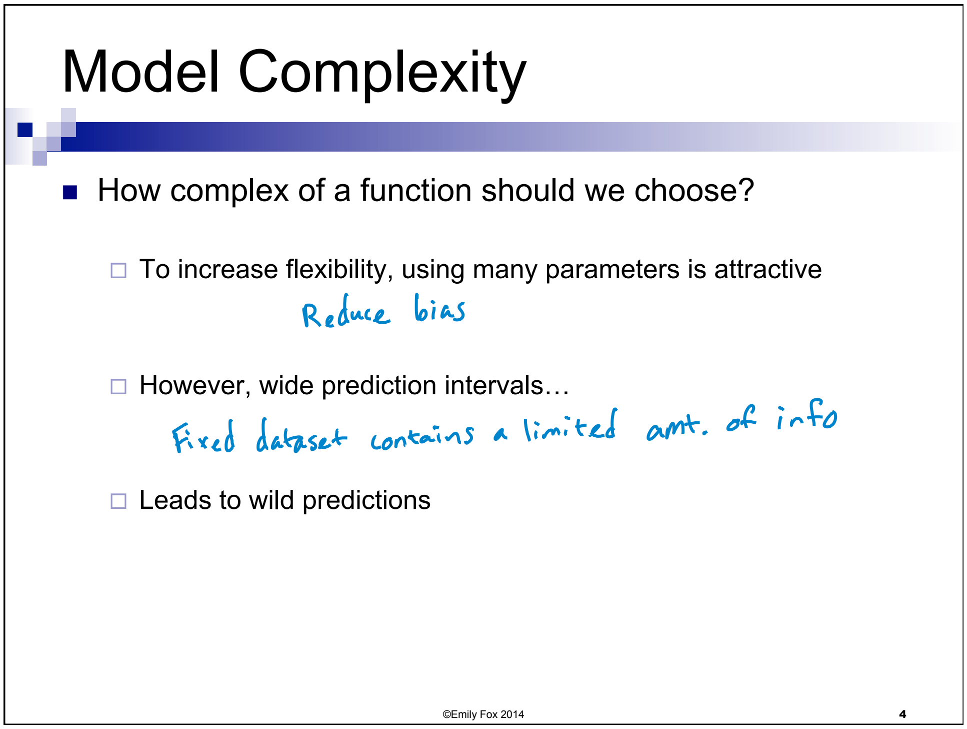
  $f(\cdot)$ is indexed by param. β

- Advantages:
  - □ Efficient estimation
  - □ Concise summarization

  e.g. LS est. of β, $\hat{\beta}_n$, leads to an est. $\hat{f}_n$ of f

- What is the right parametric form for $f(x)$?

  Should it change w/ sample size?



Model Complexity

- How complex of a function should we choose?
  - □ To increase flexibility, using many parameters is attractive

    Reduce bias

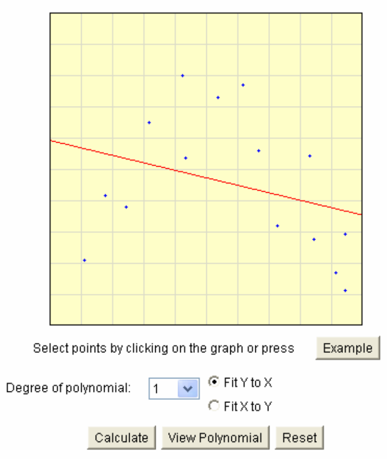
  - □ However, wide prediction intervals…

    Fixed dataset contains a limited amt. of info

  - □ Leads to wild predictions

Example: Polynomial Regression

- For added flexibility, allow for high order polynomial, right?



Select points by clicking on the graph or press

Degree of polynomial: 1

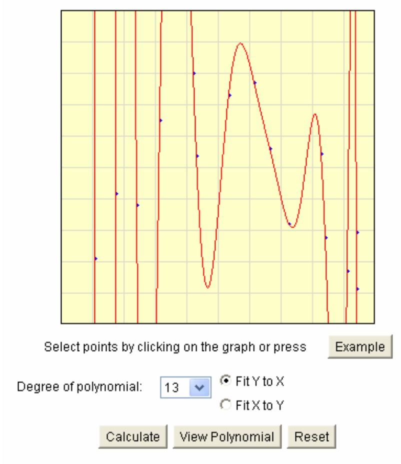
● Fit Y to X
○ Fit X to Y

$$y_i = \sum_{j=0}^{p} \beta_j x_i^j + \epsilon_i$$

Not always good to add params

Example: Polynomial Regression

- For added flexibility, allow for high order polynomial, right?



Select points by clicking on the graph or press

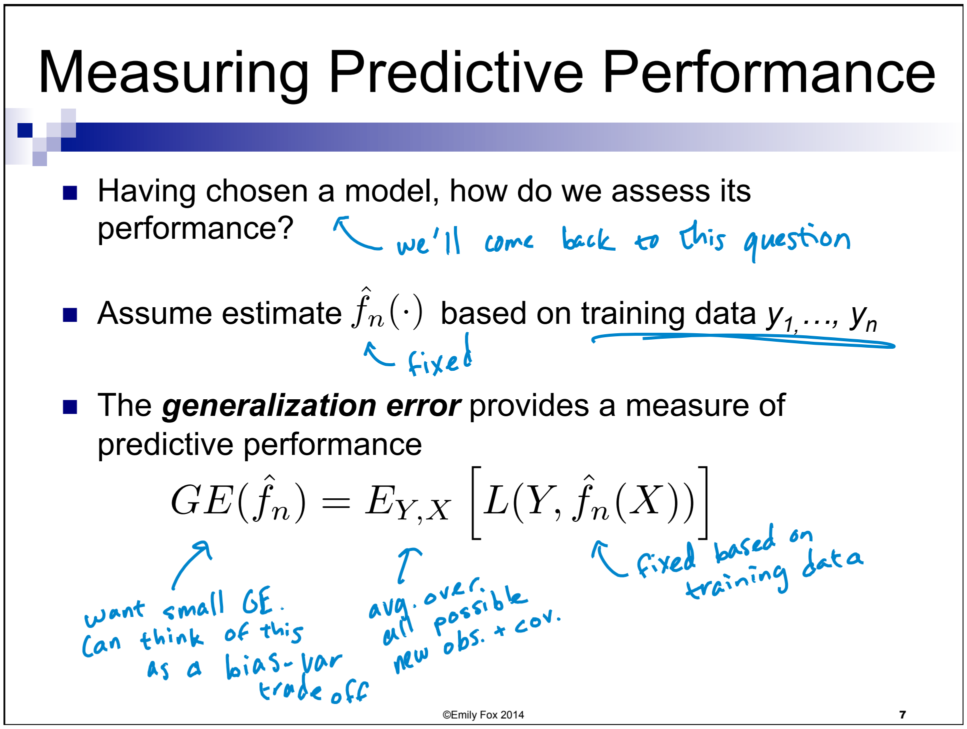
Degree of polynomial: 13

● Fit Y to X
○ Fit X to Y

sensitive to small changes in data

High order = low bias, but high var

How do we assess an estimator $\hat{f}_n$?



Measuring Predictive Performance

- Having chosen a model, how do we assess its performance? ← we'll come back to this question
- Assume estimate $\hat{f}_n(\cdot)$ based on training data $y_1, \ldots, y_n$

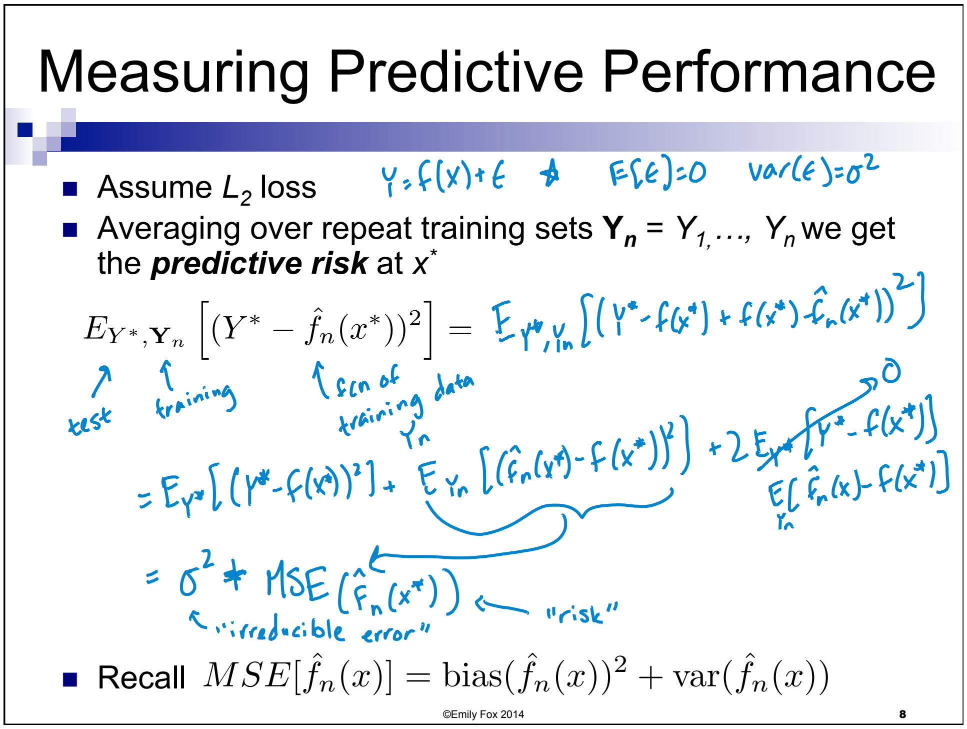
  ← fixed
- The ***generalization error*** provides a measure of predictive performance

$$GE(\hat{f}_n) = E_{Y,X}\left[L(Y, \hat{f}_n(X))\right]$$

want small GE. Can think of this as a bias-var trade off

avg. over all possible new obs. + cov.

fixed based on training data



Measuring Predictive Performance

- Assume L_2 loss $Y = f(x) + \epsilon$ * $E[\epsilon] = 0$ $\mathrm{var}(\epsilon) = \sigma^2$
- Averaging over repeat training sets $\mathbf{Y}_n = Y_1, \ldots, Y_n$ we get the ***predictive risk*** at x^*

$$E_{Y^*,\mathbf{Y}_n}\left[(Y^* - \hat{f}_n(x^*))^2\right] = E_{Y^*, Y_n}\left[(Y^* - f(x^*) + f(x^*) - \hat{f}_n(x^*))^2\right]$$

test training ← fcn of training data Y_n

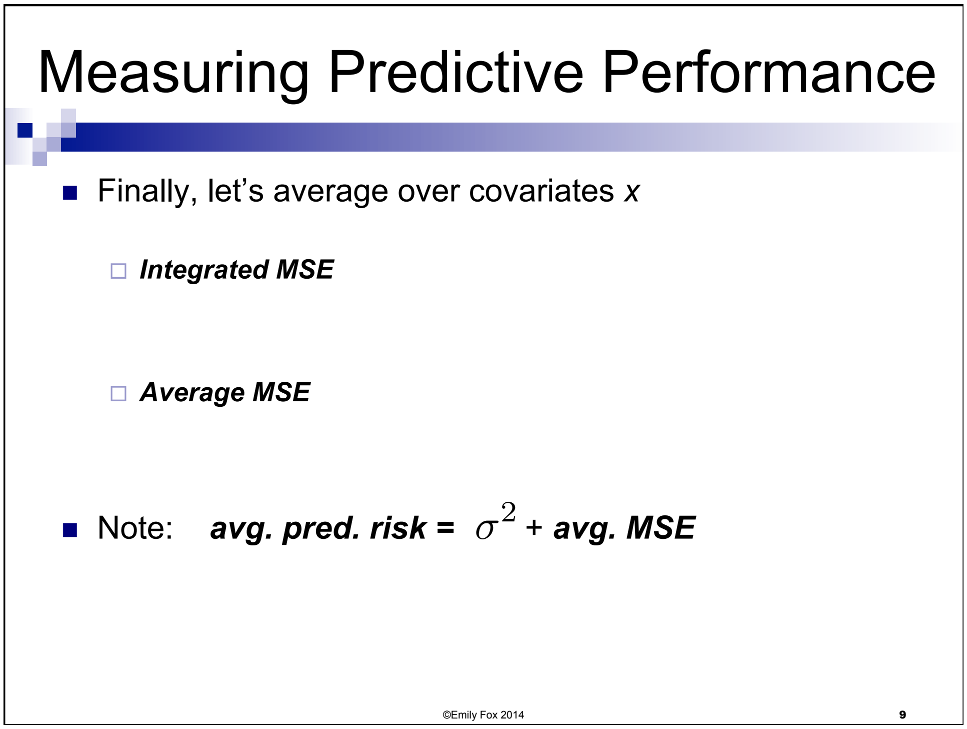
$$= E_{Y^*}\left[(Y^* - f(x^*))^2\right] + \underbrace{E_{Y_n}\left[(\hat{f}_n(x^*) - f(x^*))^2\right]} + 2\,\cancel{E_{Y^*}\left[Y^* - f(x^*)\right]} \to 0$$

$$E_{Y_n}\left[\hat{f}_n(x) - f(x^*)\right]$$

$$= \sigma^2 + MSE(\hat{f}_n(x^*))$$

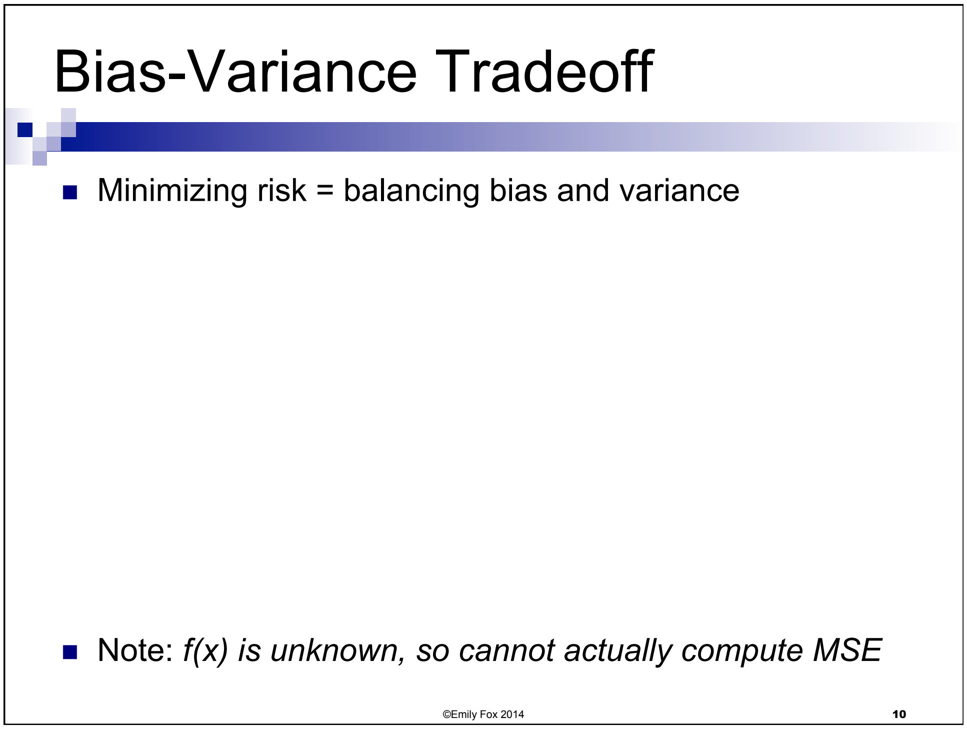
↖ "irreducible error" ← "risk"

- Recall $MSE[\hat{f}_n(x)] = \mathrm{bias}(\hat{f}_n(x))^2 + \mathrm{var}(\hat{f}_n(x))$



Measuring Predictive Performance

- Finally, let's average over covariates *x*
  - □ ***Integrated MSE***
  - □ ***Average MSE***
- Note: ***avg. pred. risk =*** σ^2 + ***avg. MSE***

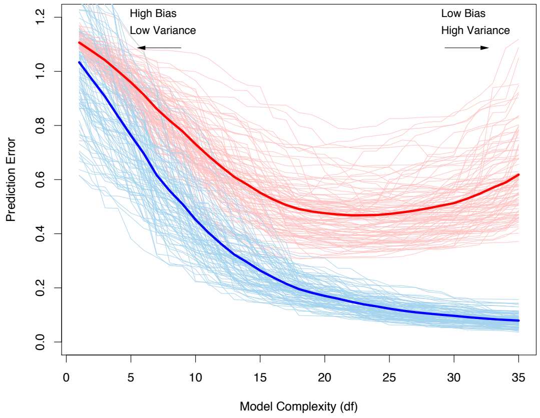


Bias-Variance Tradeoff

- Minimizing risk = balancing bias and variance
- Note: *f(x) is unknown, so cannot actually compute MSE*

In Practice…

- Minimizing risk = balancing bias and variance



From Hastie, Tibshirani, Friedman

More on Nonparam Regression

- Often framed as learning functions with a complexity penalty
  - □ Regular behavior in small neighborhoods of the input
  - □ E.g., locally linear or low-order polynomial…estimator results from averaging over these local fits

- Choice of neighborhood = strength of constraint
  - □ Large neighborhood can lead to linear fit (very restrictive) whereas small neighborhoods can lead to interpolation (no restriction)

More on Nonparam Regression

- Different restrictions lead to different nonparametric approaches
  - Roughness penalty → ***splines***
  - Weighting data locally → ***kernel methods***
  - Etc.

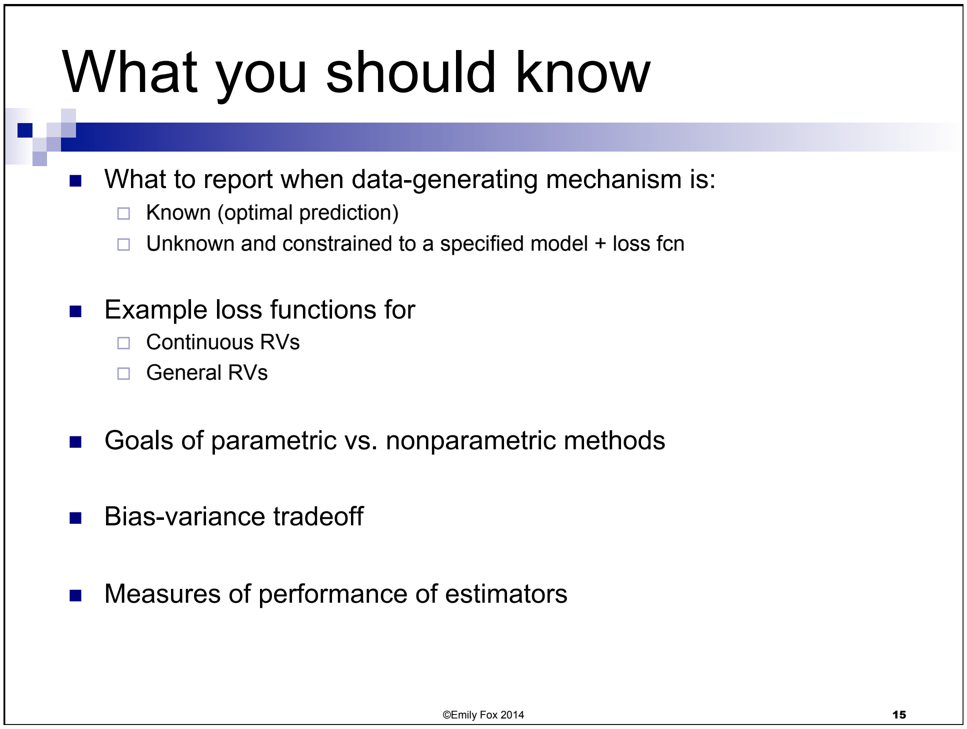
- Each method has associated ***smoothing*** or ***complexity*** param
  - Magnitude of penalty
  - Width of kernel (defining "local")
  - Number of basis functions
  - …

- Bias-variance tradeoff

- Will explore methods for choosing smoothing parameters

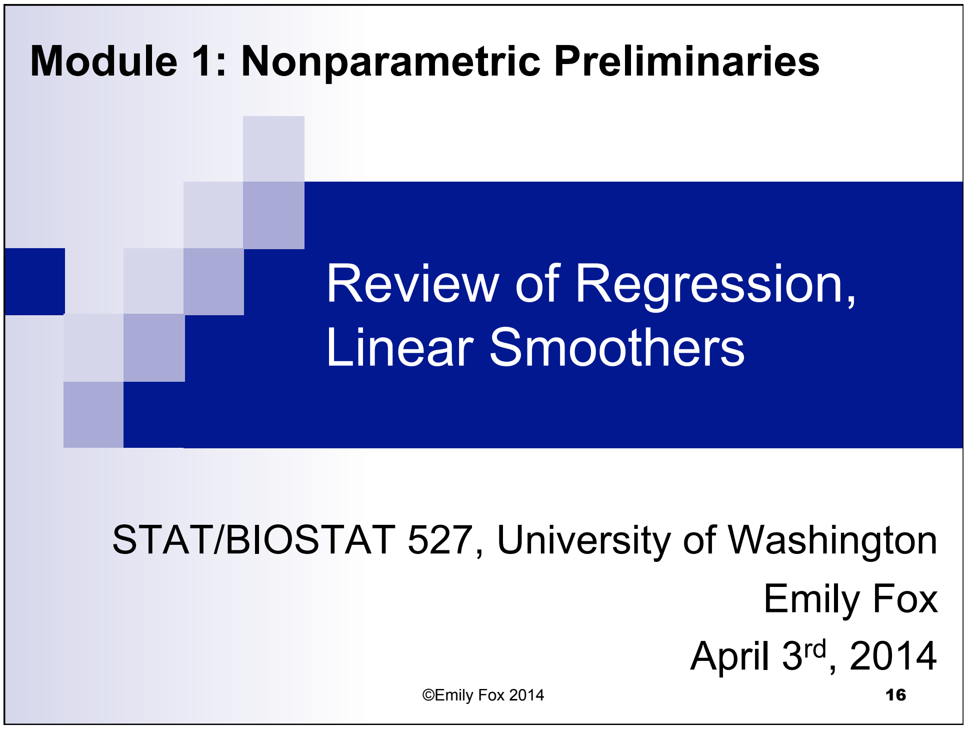
Reading

- Wakefield: 10.3-10.4
- Hastie, Tibshirani, Friedman: 7.1-7.3



What you should know

- What to report when data-generating mechanism is:
  - Known (optimal prediction)
  - Unknown and constrained to a specified model + loss fcn
- Example loss functions for
  - Continuous RVs
  - General RVs
- Goals of parametric vs. nonparametric methods
- Bias-variance tradeoff
- Measures of performance of estimators



Module 1: Nonparametric Preliminaries

Review of Regression, Linear Smoothers

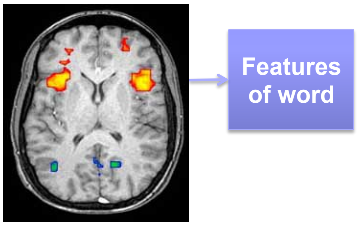
STAT/BIOSTAT 527, University of Washington

Emily Fox

April 3rd, 2014

fMRI Prediction Subtask

- **Goal:** Predict semantic features from fMRI image

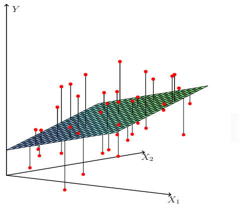


Features of word

Linear Regression – *review*

- Model:
- ***Design matrix***:
- Rewrite in matrix form:

Least Squares



- Least squares estimation:
  - □ Minimize ***residual sum of squares***
  - □ Take gradient and set = 0
- In Gaussian case, LS est. = maximum likelihood est.

Fitted Values

- ***Fitted values***
- Number of parameters
- For any *x*, we can write

Linear Smoothers

- Definition:
  $\hat{f}_n$ of f is a ***linear smoother*** if, for each *x*, there exists
  $$\ell(x) = (\ell_1(x), \ldots, \ell_n(x))^T$$
  such that
- Matrix form
  - Fitted values
  - Smoothing or “hat” matrix
- Effective degrees of freedom:

Linear Smoothers

- Note 1:

  A linear smoother does ***not*** imply that $f(x)$ is linear in *x*
- Note 2:

  If $Y_i = c$ for all *i*, then $\hat{f}_n(x) = c$ for all *x*

Module 1: Nonparametric Preliminaries

Overfitting, Ridge Regression, LASSO

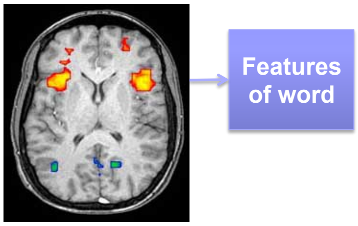
STAT/BIOSTAT 527, University of Washington

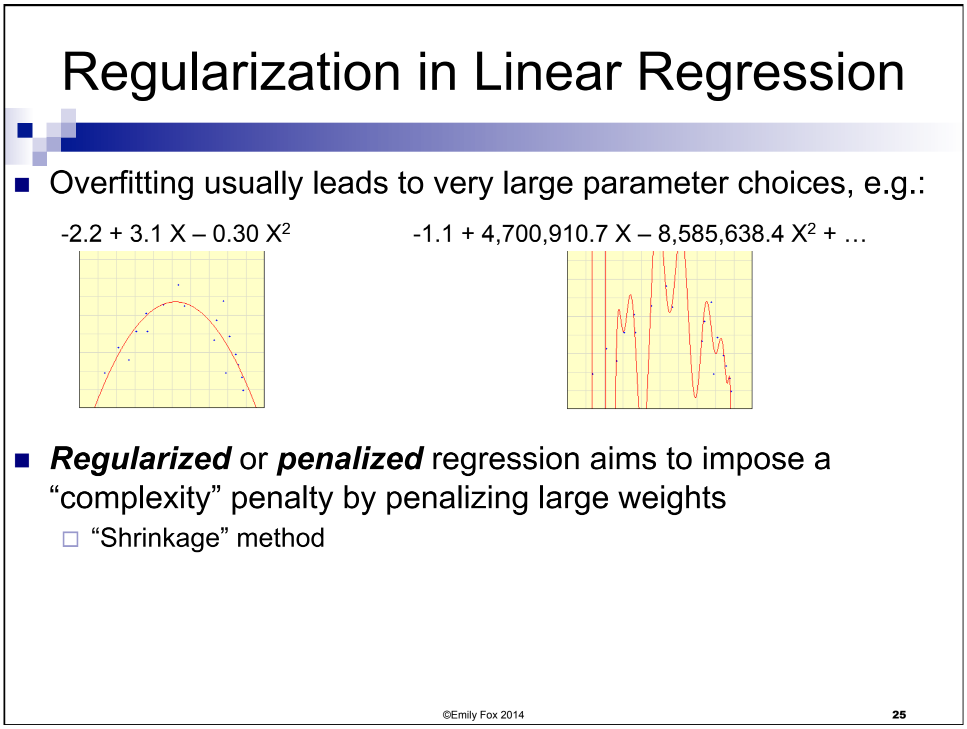
Emily Fox

April 3rd, 2014

fMRI Prediction Subtask

- **Goal:** Predict semantic features from fMRI image



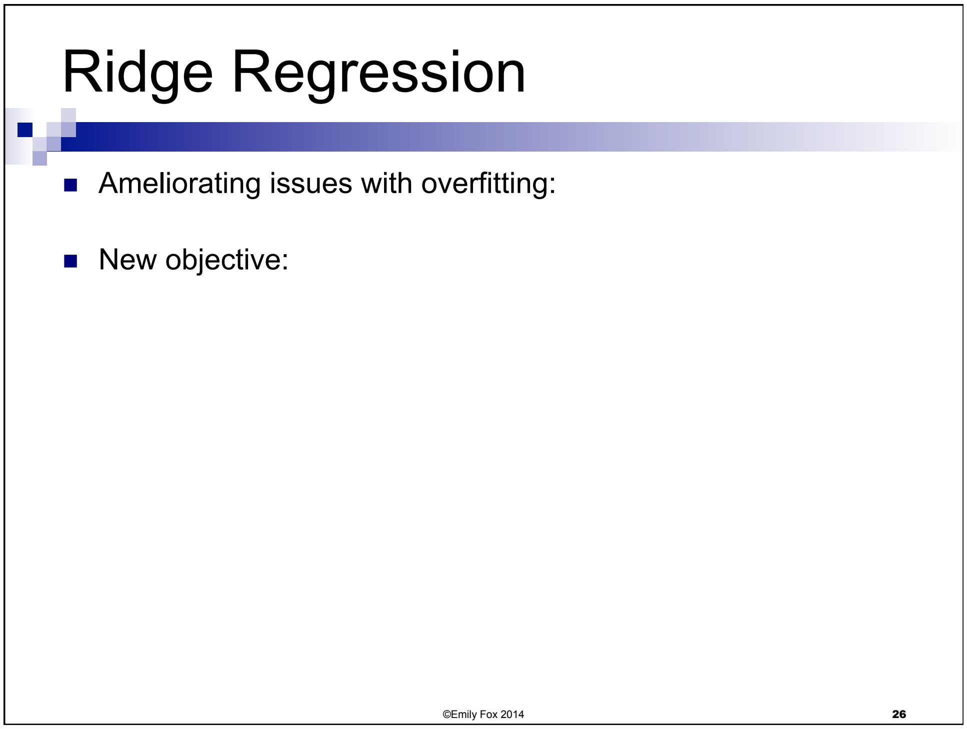


Regularization in Linear Regression

- Overfitting usually leads to very large parameter choices, e.g.:

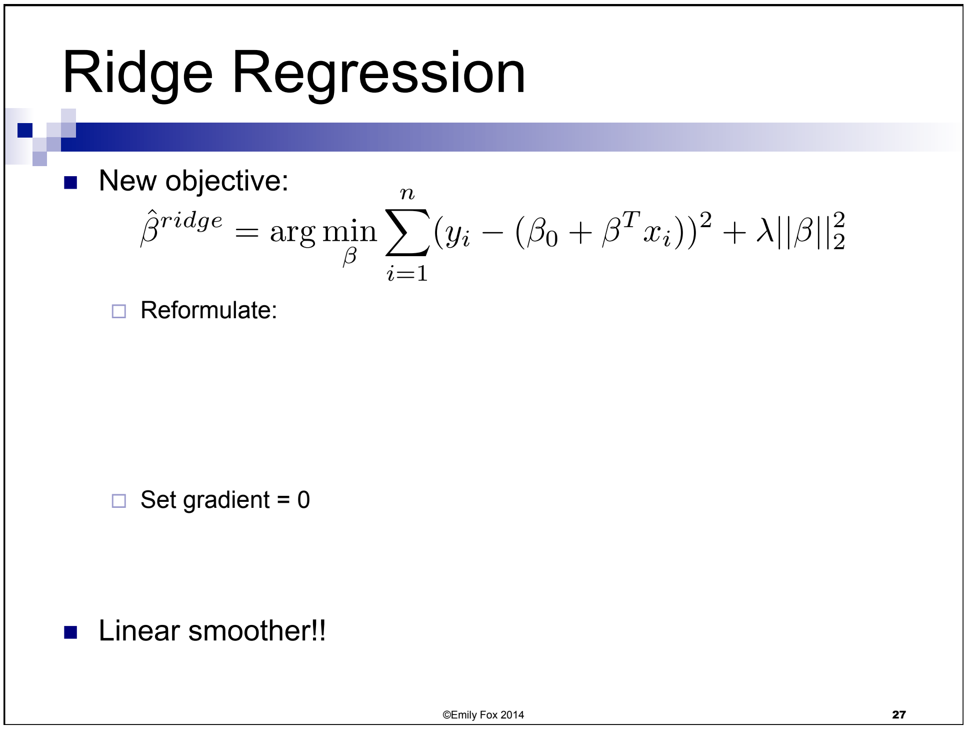
$-2.2 + 3.1\ X - 0.30\ X^2$ $\quad\quad$ $-1.1 + 4{,}700{,}910.7\ X - 8{,}585{,}638.4\ X^2 + \ldots$

- ***Regularized*** or ***penalized*** regression aims to impose a "complexity" penalty by penalizing large weights
  - "Shrinkage" method



Ridge Regression

- Ameliorating issues with overfitting:
- New objective:

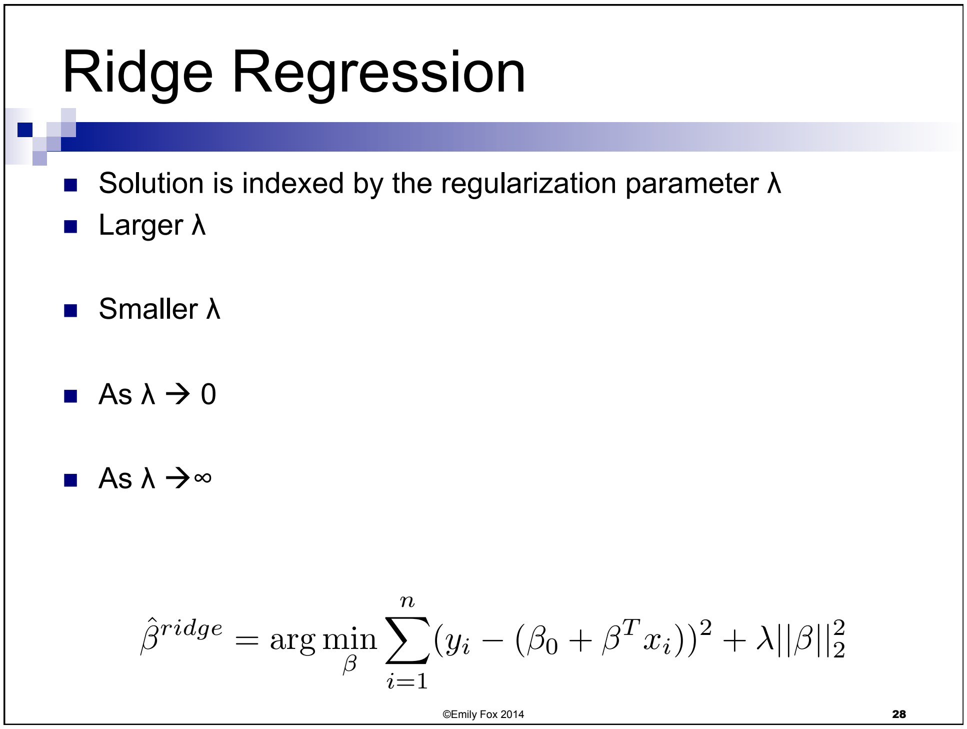


Ridge Regression

- New objective:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta^T x_i))^2 + \lambda ||\beta||_2^2$$

  - □ Reformulate:
  - □ Set gradient = 0
- Linear smoother!!



Ridge Regression

- Solution is indexed by the regularization parameter λ
- Larger λ
- Smaller λ
- As λ → 0
- As λ →∞

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta^T x_i))^2 + \lambda ||\beta||_2^2$$

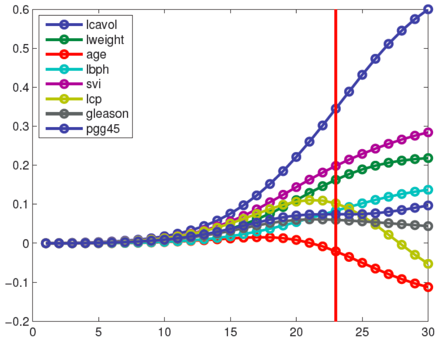
Shrinkage Properties

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

- If orthogonal covariates $X^T X = I$

- Effective degrees of freedom:

Ridge Coefficient Path



From
Kevin Murphy
textbook

- Typical approach: select λ using cross validation

A Bayesian Formulation

- Consider a model with likelihood

$$y_i \mid \beta \sim N(\beta_0 + x_i^T \beta, \sigma^2)$$

and prior

$$\beta \sim N\left(0, \frac{\sigma^2}{\lambda} I_p\right)$$

- For large λ

- The posterior is

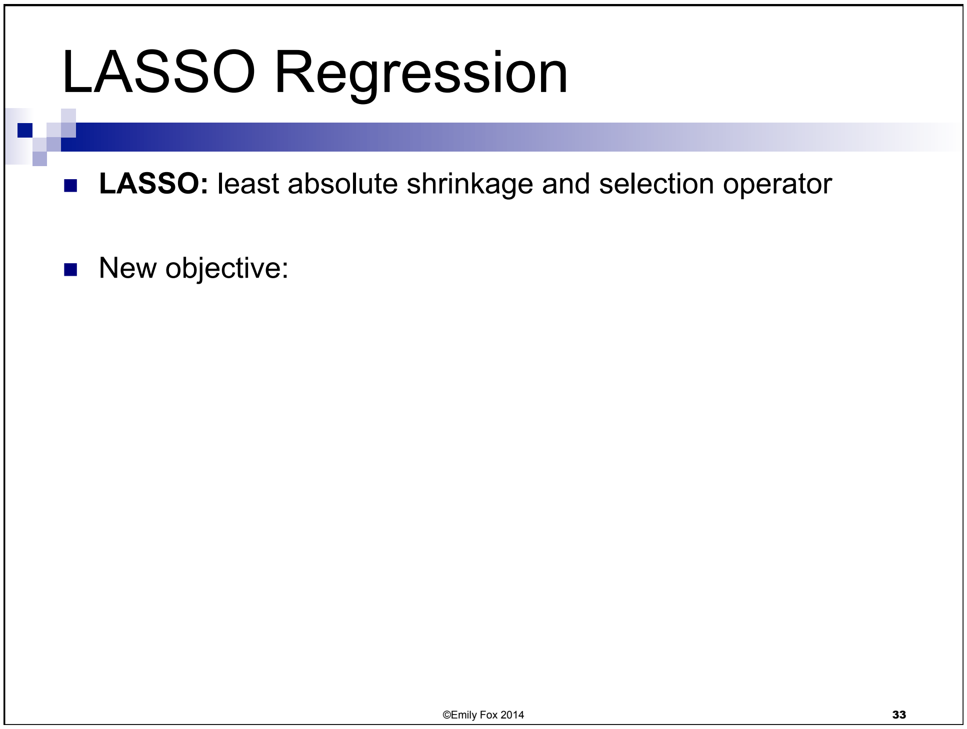
$$\beta \mid y \sim N\left(\hat{\beta}^{ridge}, \sigma^2 (X^T X + \lambda I)^{-1} X^T X \sigma^2 (X^T X + \lambda I)^{-1}\right)$$

Variable Selection

- Ridge regression: Penalizes large weights

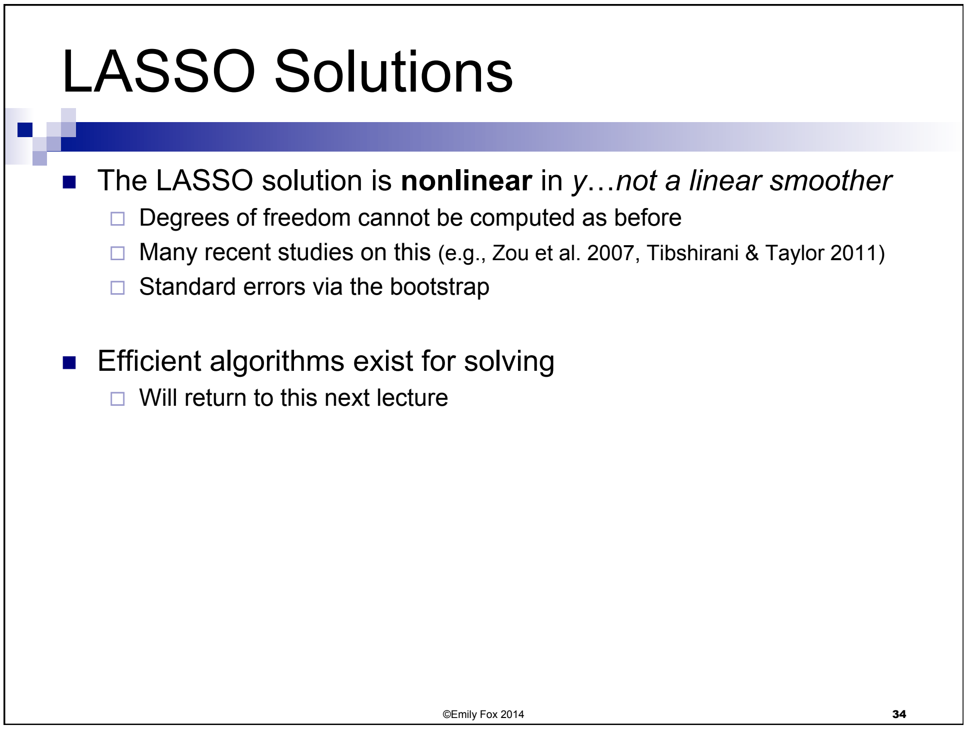
- What if we want to perform “feature selection”?
  - □ E.g., Which regions of the brain are important for word prediction?
  - □ Can’t simply choose predictors with largest coefficients in ridge solution
  - □ Computationally impossible to perform “all subsets” regression
  - □ Stepwise procedures are sensitive to data perturbations and often include features with negligible improvement in fit

- Try new penalty: Penalize non-zero weights
  - □ Penalty:
  - □ Leads to sparse solutions
  - □ Just like ridge regression, solution is indexed by a continuous param λ



LASSO Regression

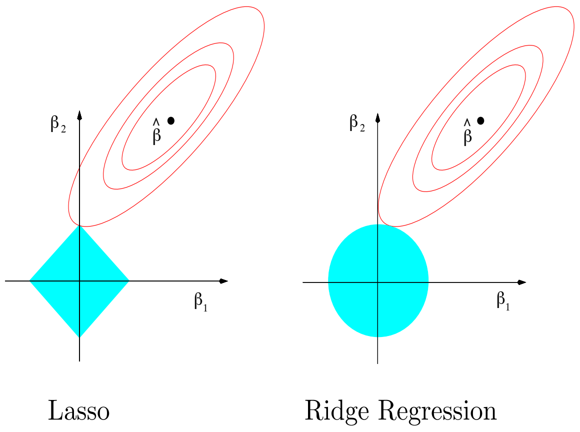
- **LASSO:** least absolute shrinkage and selection operator
- New objective:



LASSO Solutions

- The LASSO solution is **nonlinear** in *y…not a linear smoother*
  - □ Degrees of freedom cannot be computed as before
  - □ Many recent studies on this (e.g., Zou et al. 2007, Tibshirani & Taylor 2011)
  - □ Standard errors via the bootstrap
- Efficient algorithms exist for solving
  - □ Will return to this next lecture

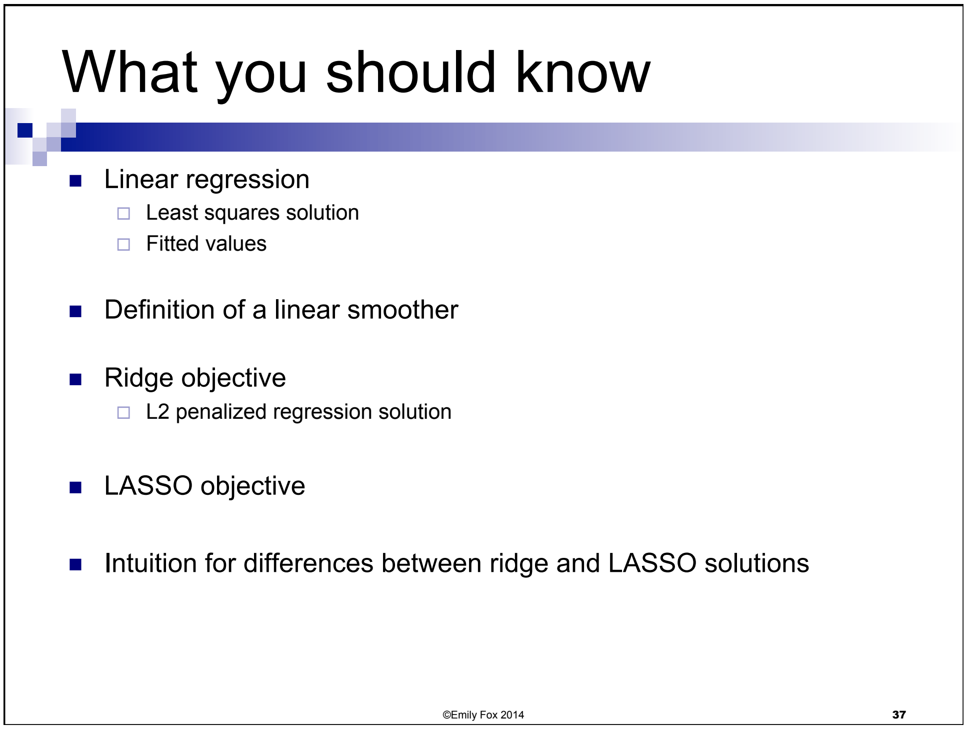
Geometric Intuition for Sparsity



From Rob Tibshirani slides

Reading

- Hastie, Tibshirani, Friedman: 3.2 (up to 3.2.3), 3.4
- Wasserman: 5.2



What you should know

- Linear regression
  - □ Least squares solution
  - □ Fitted values
- Definition of a linear smoother
- Ridge objective
  - □ L2 penalized regression solution
- LASSO objective
- Intuition for differences between ridge and LASSO solutions