Regularized Regression

Ridge and The Lasso

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

A Note

- Not just for continuous outcomes – Ridge and Lasso have extensions for logistic (and other) regression models.

- In the corresponding code from R glmnet package, simply add the option family="binomial" for a classification task.

Regularization and Overfitting

The bias-variance tradeoff revisited

Overfitting

- Models with too many variables will overfit the training data.

IF you want a linear model, but:

- Leaving variables out is not an option.
- **You find it too difficult to determine which variables to leave out.**
- You need many variables to model a significant portion of signal.
- You have more variables than observations.
- You want superior predictions on out-of-sample data.

Then Regularized Regression is your best bet.

Bias-Variance Tradeoff

- The mean-squared error of a model on out-of-sample test data can be decomposed into three terms:

$$E(y - f(x))^2 = Var(f(x)) + [Bias(f(x))]^2 + Var(\epsilon)$$

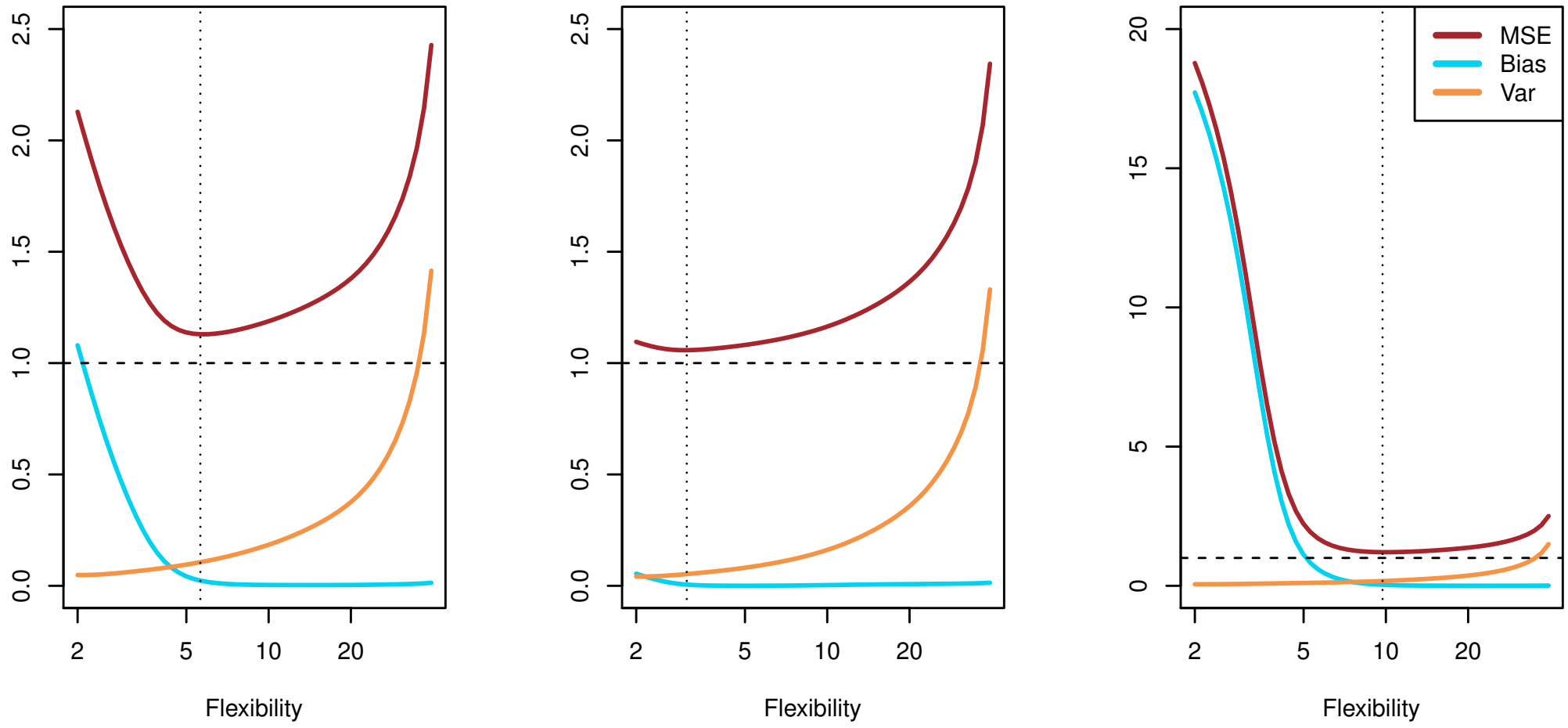
The **variance** of the estimates when the model is created on different training sets

The squared **bias** of the estimates (The squared difference between the average estimate over different training sets and the actual target value)

The irreducible error (can't model this)

Bias-Variance Tradeoff Illustrated

(For 3 different data sets)



(x-axis represents model complexity)

Regularization

- In Machine Learning, a common tool is "regularization"
- Regularization adds a penalty term to the objective function of a model that penalizes model complexity.
- Also called **parameter shrinkage**
- Regularization has been shown to trade the introduction of small amounts of bias for a reduction in large amounts of variance.

Strong Feelings…

”If you’re using regression without regularization, you have to be very special…”

– Owen Zhang (Kaggle rank 3, Chief Product Officer at Data Robot)

What if I told you…

- You could just go ahead and keep ALL of your variables in the model
- Without overfitting the training data
- Ending with a complex but totally generalizable model

Ridge Regression

a.k.a.
Tikhonov Regularization
L_2 Regularization
Weight-decay
(Tikhonov **1943**)

Ridge Regression

- Ridge regression is a biased regression technique (like PCR)
- Parameter estimates tend to have lower variance than OLS estimates, but are biased
- Often proposed as a 'solution' for multicollinearity when estimating parameters.

- Theoretically shown to trade large amounts of variance for minimal amounts of bias

Ridge Regression

- OLS minimizes the sum of squared error:
  - OLS Objective function: $f_{OLS}(x) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Ridge regression adds a penalty for the parameters in the model:
  - Ridge Objective function: $f_{RIDGE}(x) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$

- **FIRST STEP IS TO STANDARDIZE YOUR DATA!**
- **SAS will do this FOR YOU when you include the ridge = option.**

λ - The Regularization Parameter

- The larger the value of λ, the more bias is introduced into the model.
- At very large values, all the parameters would be forced to zero.
- At very small values, the penalty term would have no effect.

- Many ways to tune this parameter!

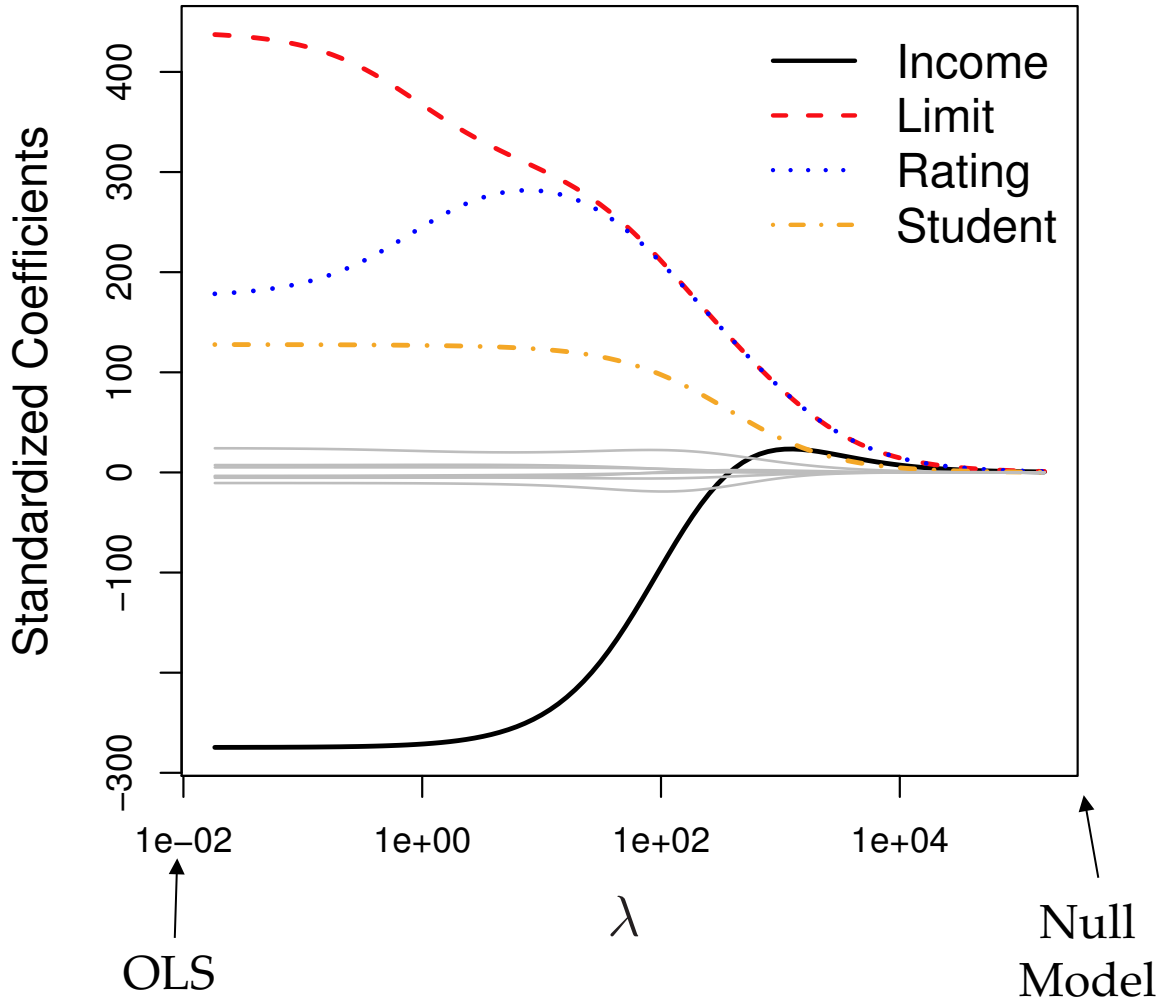
λ - The Regularization Parameter

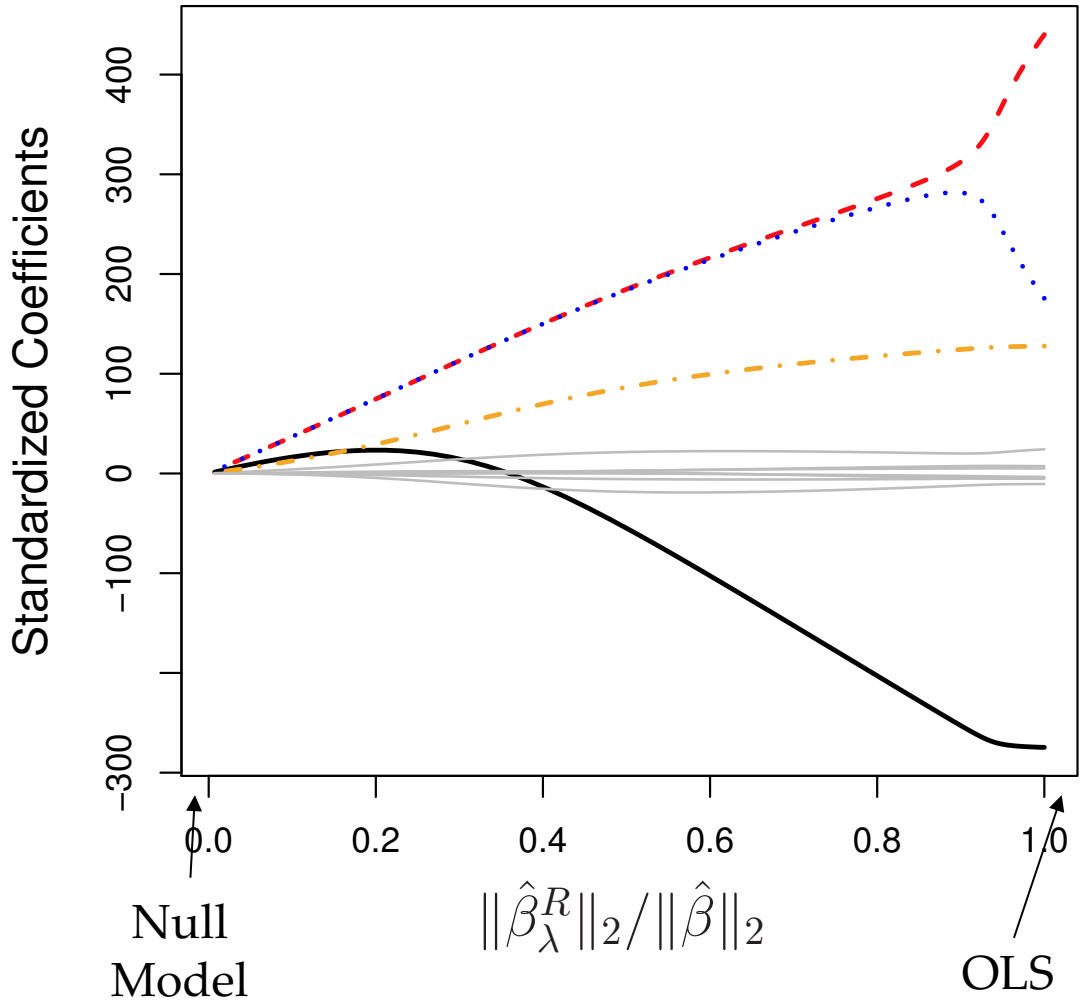
- Fixed point estimate for λ (1975 – Hoerl, Kennard, Baldwin)

$$\lambda = \frac{p\,\widehat{\sigma^2}}{\sum_{i=1}^{p} \hat{\beta}_{OLS}{}^2}$$

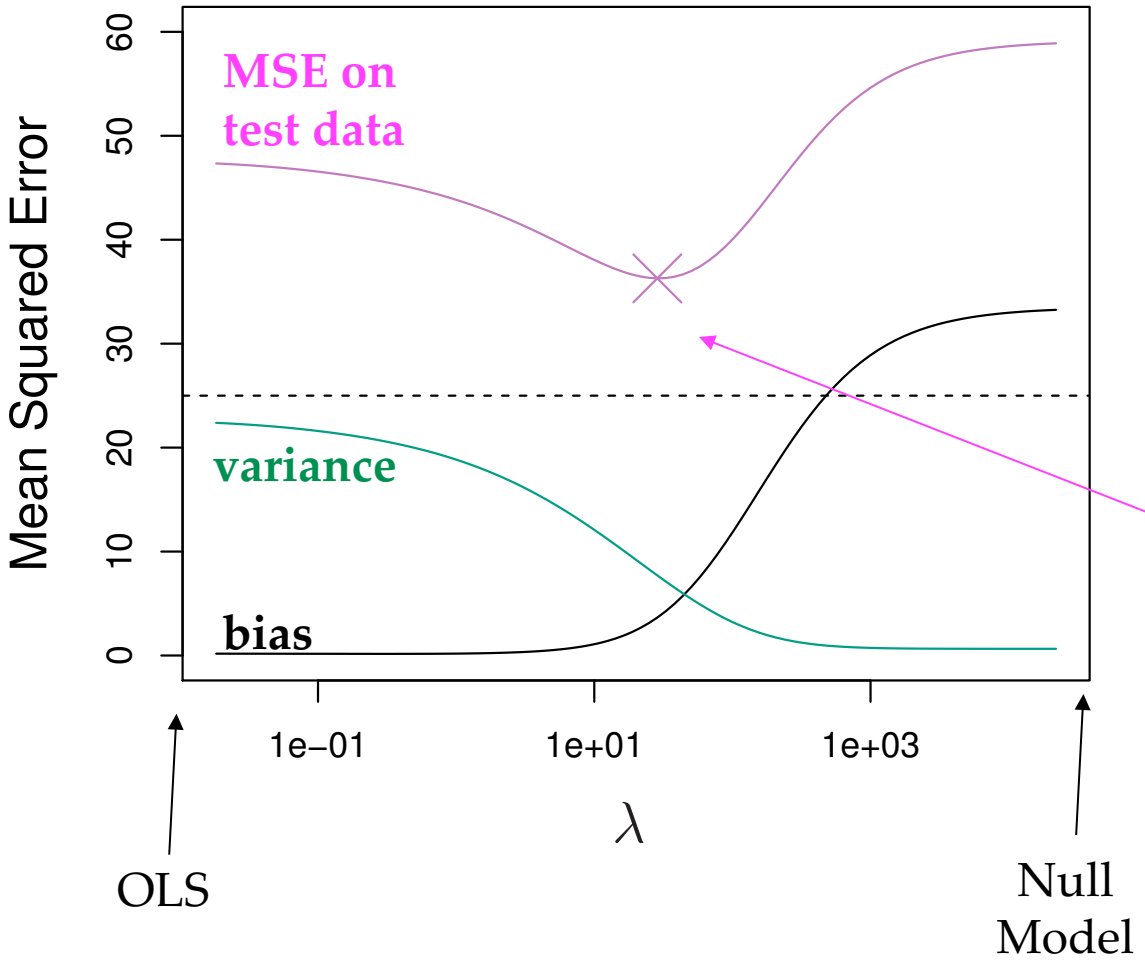
- Iterative Method (1976 – Hoerl, Kennard)
- Ridge Trace Method
  - Plot of many different estimates of $\widehat{\beta_i}$ across a series of values of λ.
  - Use the plot to approximate when the estimates become stable and have "sensible" signs.
  - No objective basis, heavily criticized by many
- Optimization using cross validation standard practice in ML community
  - Chose λ that provides the minimum average error on K-fold cross validation.

How λ affects parameters





How λ affects bias/variance/MSE



- When λ small, no penalty
  - high variance
  - no bias
  - high MSE on test data
- When λ big, null model
  - high bias
  - no variance
  - high MSE on test data
- Sweet spot
  - minimizes MSE on test data
  - introduces small bias
  - substantially reduces variance

Ridge Regression

- Will allow you to keep all variables in your model without the problem of overfitting.
- Works best in situations where least squares estimates have high variance
- Will not create many zero parameter estimates, so all of the input variables likely to stay in the model.

The LASSO

• • •

a.k.a

L_1 Regularization

(Tibshirani **1996**)

Penalties for Model Selection

- In recent years, stepwise selection techniques for variable selection have come under fire.
- Alternative methods, such as "The LASSO" have been proposed and have soared in popularity.

Drawbacks to Stepwise Selection

- Bias in parameter estimation
  - Standard errors biased toward zero
  - p-values biased toward zero
  - Parameter estimates biased away from zero
  - R-Squared biased upwards
- F and Chi-Square tests don't have the desired distribution
- Resulting models are complex with exacerbated collinearity problems
- Inconsistencies among model selection algorithms
- An inherent problem with multiple hypothesis testing
- An inappropriate focus or reliance on a single best model

(MJ Whittingham et al – 2006, Harrell – 2010, Flom & Cassell – 2007)

Analogy for Stepwise

Flom and Cassell (2007) write:

> *"In Stepwise Regression, this assumption* [of independent hypothesis tests] *is grossly violated in ways that are difficult to determine.*
>
> *For example, if you toss a coin ten times and get ten heads, then you are pretty sure that something weird is going on. You can quantify exactly how unlikely such an event is…*
>
> *If you have 10 people each toss a coin ten times, and one of them gets 10 heads, you are less suspicious, but you can still quantify the likelihood.*
>
> *But if you have a bunch of friends (you don't count them) toss coins some number of times (they don't tell you how many) and someone gets 10 heads in a row, you don't even know how suspicious to be. That's stepwise."*

LASSO Regression

- OLS minimizes the sum of squared error:
  - OLS Objective function: $\mathrm{f_{OLS}(x)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- LASSO regression adds a penalty for the parameters in the model:
  - LASSO Objective function: $\mathrm{f_{LASSO}(x)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$

- **FIRST STEP IS TO STANDARDIZE YOUR DATA!**

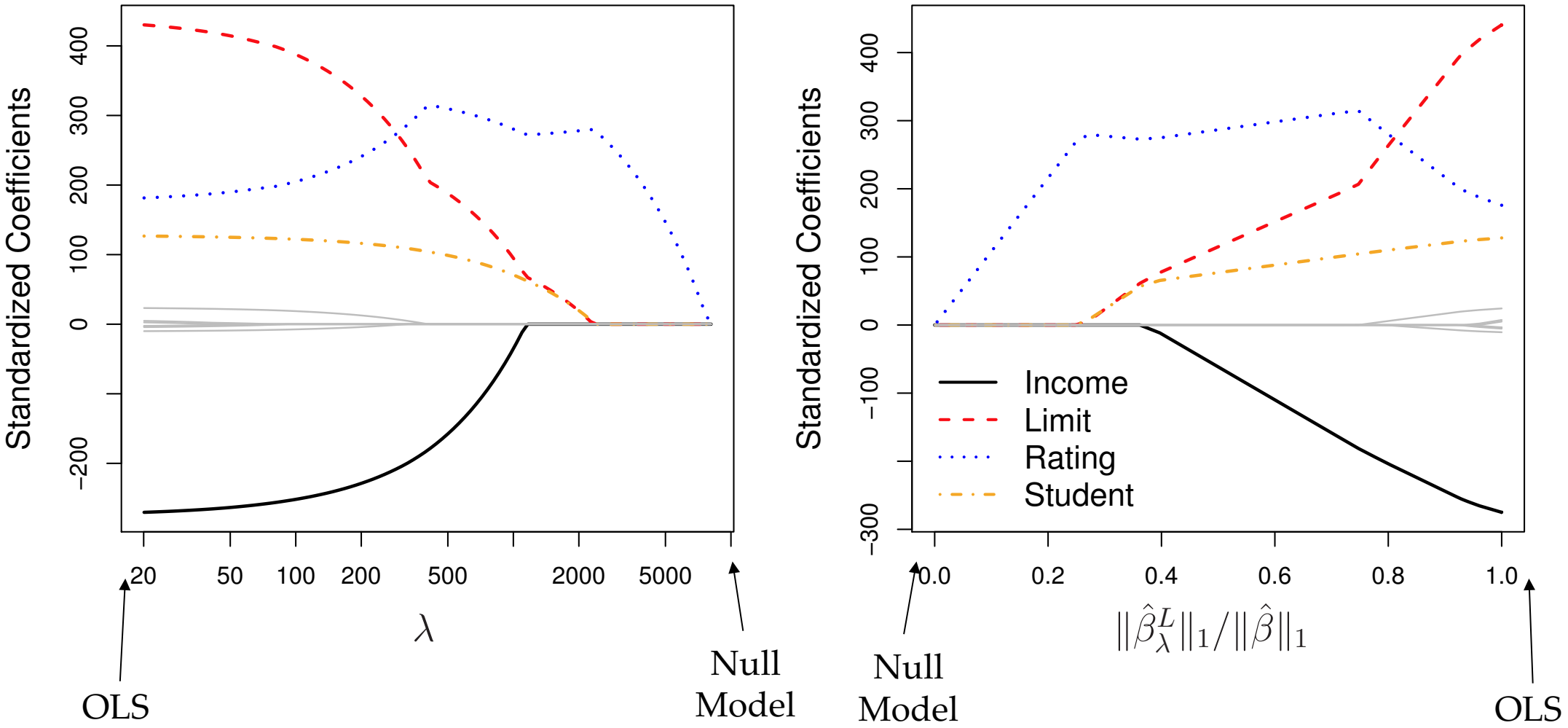
LASSO Regression

- The LASSO penalty has the added benefit that it typically causes many of the parameter estimates to tend toward zero.

=> The LASSO produces **sparse solutions**

- This implies automated variable selection

How λ affects Parameters



LASSO Regression

- Very common when the number of variables is overwhelming for stepwise selection (particularly in text)

- Generally implemented through Least Angle Regression (LARS) algorithm (Efron et al 2004)
  - glmnet package in R
  - lars package in R
  - LARS node in SAS EM
  - proc glmselect option selection=LASSO

Predicting Salary of Baseball Players

An Example in R

Stepwise Selection Using Validation

The purpose here is to choose a level of model complexity, p= the number of parameters.

1. Run the stepwise selection algorithm on training data. For all possible number of variables, p, find the chosen model.
2. Compare the p models found in step 1 on validation data and record the MSE.
3. Pick the "optimal" number of parameters p* as the one that minimized the MSE on validation data.
4. Now you've validated your modelling process. Re-run the stepwise selection on the entire data to choose p* parameters.

Yes, They may be different when you use all the data! That's ok. You've validated the procedure!

Stepwise Selection Using Validation

- In the Hitters example, we found that forward selection chose a model with 10 parameters and backward selection chose a model with 7 parameters.
- The validation MSE on both models was ~the same.
- Choose the simpler model.

Ridge Regression

- The number of parameters used in the stepwise selection models didn't agree, nor did the actual variables used.
- If simplicity of the model is not as important as generalizability, we can consider ridge regression.
- Why?
  - Skip the agony of choosing predictors. Use them all, and shrink parameters to control for overfitting.
  - Ridge regression generally yields better predictions than OLS through a better bias-variance compromise.
  - Works especially well in the presence of severe multicollinearity in predictor variables.

The LASSO

- If simplicity of the model IS desired, and we decide that we're not comfortable with the stepwise selection procedure…

  (In the case of many variables (~50-100) we probably shouldn't be comfortable with such a procedure)

- Parameter shrinkage methods like the LASSO have proven to work better in light of the bias-variance trade-off.

Lambda.min vs Lambda.1se

- Lambda.min
  - The value of lambda that provides the minimum average MSE on cross validation
- Lambda.1se
  - The value of lambda that provides the simplest model but still provides MSE within 1 standard error of the minimum of cross validation

- Sometimes lambda.min provides a model that still has too much variance and lambda.1se provides a slightly more stable model with less variance

Extra Bits for Self-Study

Ridge Trace Method in SAS

```
proc reg data=frenchstd outvif outest=B
                        ridge=0 to 0.08 by 0.002;
     model Import = DoProd Stock Consum / vif;
run;
quit;

proc reg data=frenchstd outvif outseb outest=B
                        ridge=0.04;
     model Import = DoProd Stock Consum / vif;
run;
quit;

proc print data=B;
run;
```

ElasticNet Criteria

(Zou & Hastie 2005)

- The ElasticNet Criteria combines the L_1 penalty and the L_2 penalty to achieve both the parameter shrinkage of ridge regression and the sparsity feature of the LASSO.

$$\mathrm{f}_{ELASTIC}(\mathrm{x}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2$$

- Works better than LASSO when some of your input variables are highly correlated
- In R glmnet package, simply set 0<alpha<1.
- Alpha closer to 0 emphasizes ridge regression
- Alpha closer to 1 emphasizes LASSO.