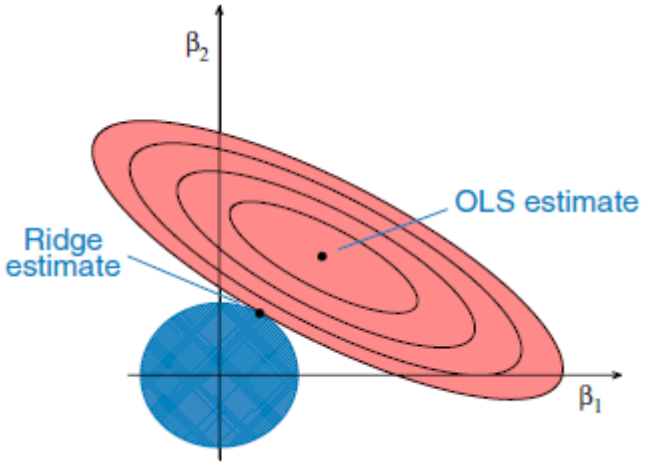
Modern Regression Approaches

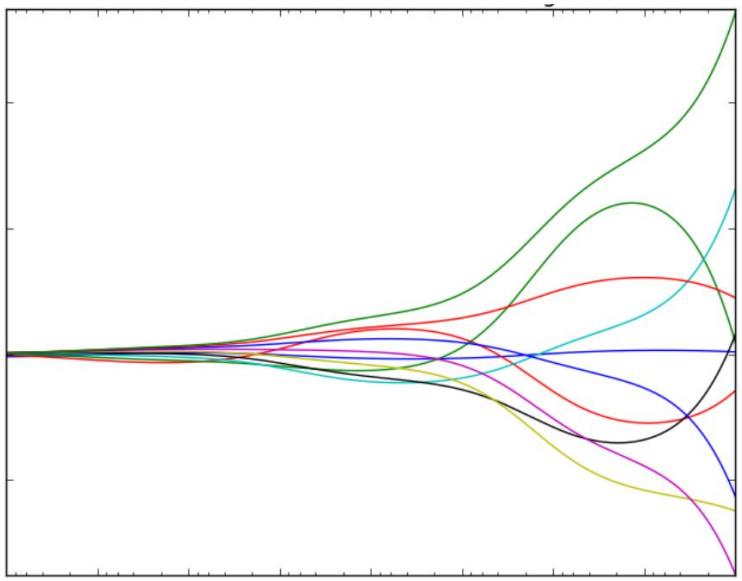
Presented by: Derek Kane

Overview of Topics

- Advancements with Regression
- Ridge Regression
- Lasso
- Elastic Net
- Practical Example
  - Prostate Cancer



Introduction to Regression Analysis

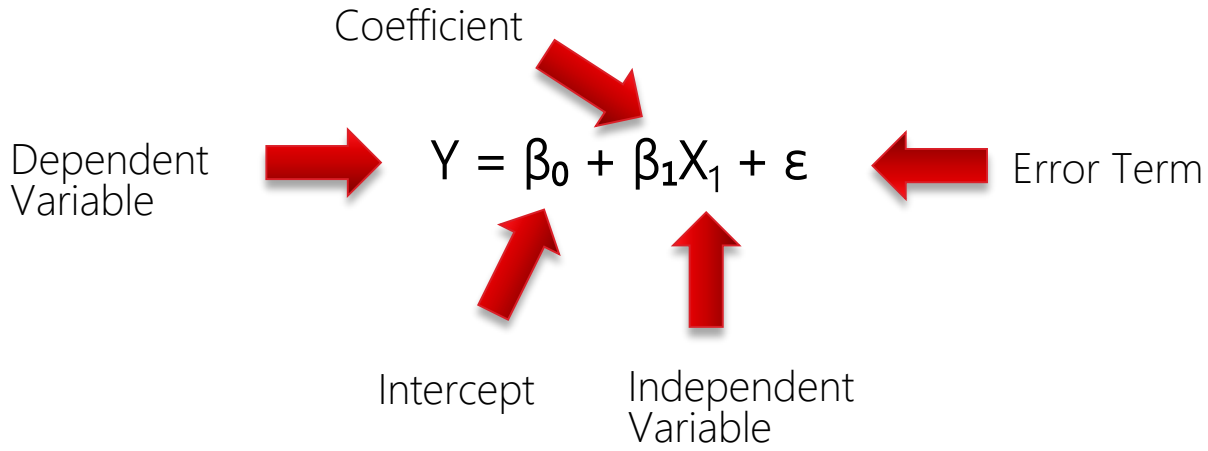


- If we continue to draw from OLS as our only approach to linear regression techniques, methodologically speaking, we are still within the late 1800's and early 1900's timeframe.
- With advancements in computing technology, regression analysis can be calculated using a variety of different statistical techniques which has lead to the development of new tools and methods.
- The techniques we will discuss today will bring us to date with advancements in regression analysis.
- In modern data analysis, we will often find data with a very high number of independent variables and we need better regression techniques to handle this high-dimensional modeling.

Review of Linear Regression Analysis

Simple Linear Regression Formula

- The simple regression model can be represented as follows:



- The β_0 represents the Y intercept value, the coefficient β_1 represents the slope of the line, the X_1 is an independent variable and ε is the error term. The error term is the value needed to correct for a prediction error between the observed and predicted value.

Review of Linear Regression Analysis

Simple Linear Regression Formula

- The output of a regression analysis will produce a coefficient table similar to the one below.

Coefficients				
Term	**Coefficient**	**Standard Error**	**T Value**	**Pr > |t|**
Intercept	-114.326	17.4425	-6.55444	0.03
Height	106.505	11.55	9.22117	0.001

- This table shows that the intercept is -114.326 and the Height coefficient is 106.505 +/- 11.55.
- This can be interpreted as for each unit increase in X, we can expect that Y will increase by 106.5
- Also, the T value and Pr > |t| indicate that these variables are statistically significant at the 0.05 level and can be included in the model.

Review of Linear Regression Analysis

Multiple Linear Regression Formula

- A multiple linear regression is essentially the same as a simple linear regression except that there can be multiple coefficients and independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

- The interpretation of the coefficient is slightly different than in a simple linear regression. Using the table below the interpretation can be thought of:

Coefficients				
Term	**Coefficient**	**Standard Error**	**T Value**	**Pr > |t|**
Intercept	-114.326	17.4425	-6.55444	0.03
Height	106.505	11.55	9.22117	0.001
Width	94.56	8.345	5.6612	0.048

- For each 1 unit change in Width, increases Y by 94.56. This is while holding all other coefficients constant.

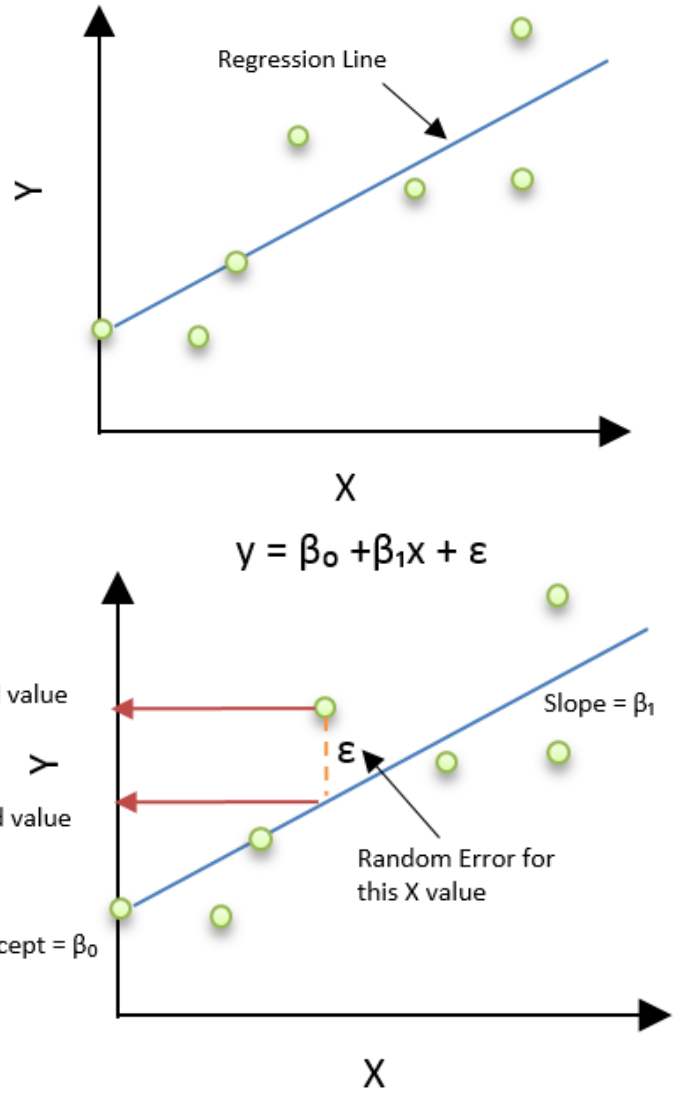
Ordinary Least Squares

What is Ordinary Least Squares or OLS?

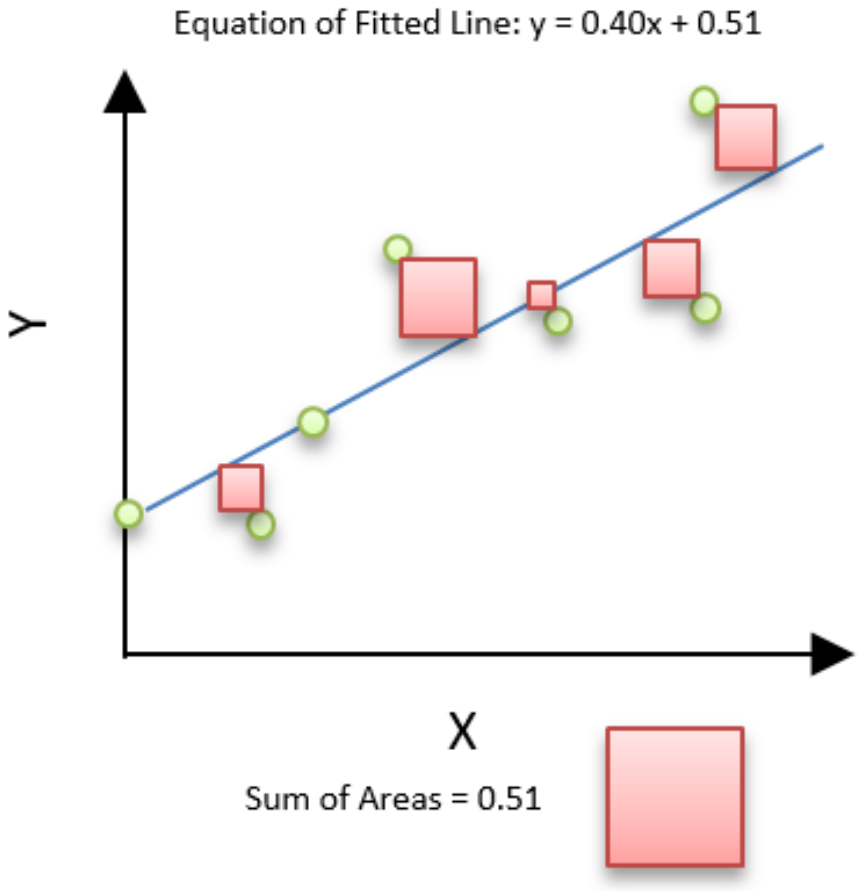
- In statistics, ordinary least squares (OLS) or linear least squares is a method for estimating the unknown parameters in a linear regression model.

- The goal of OLS is to minimize the differences between the observed responses in some arbitrary dataset and the responses predicted by the linear approximation of the data.

$$y_n = \sum_{i=0}^{k} \beta_i x_{ni} + \varepsilon_n$$



Ordinary Least Squares



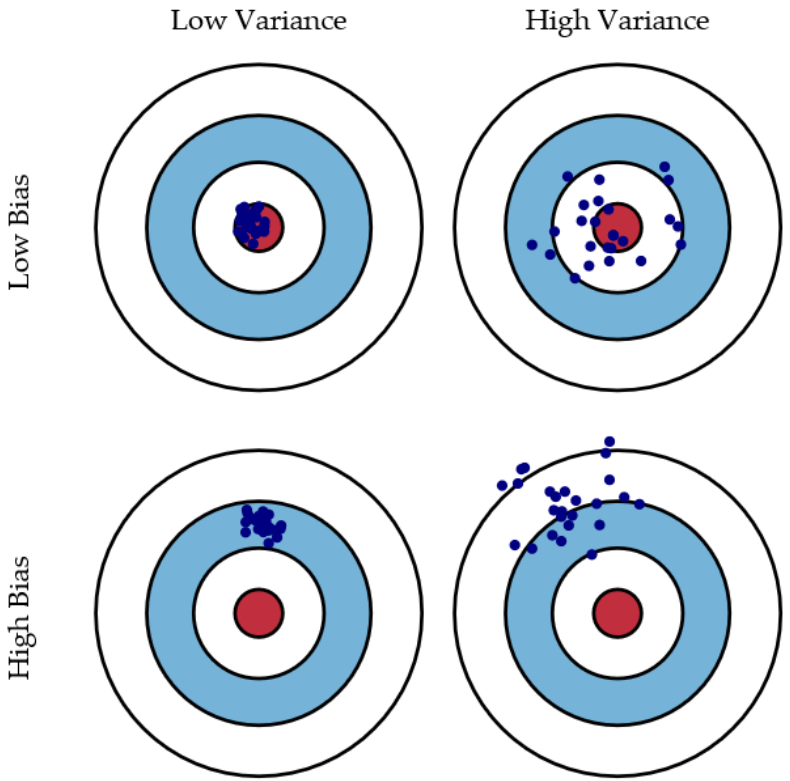
- Visually this is seen as the sum of the vertical distances between each data point in the set and the corresponding point on the regression line.
- The smaller the differences (square size), the better the model fits the data.

Understanding the Error

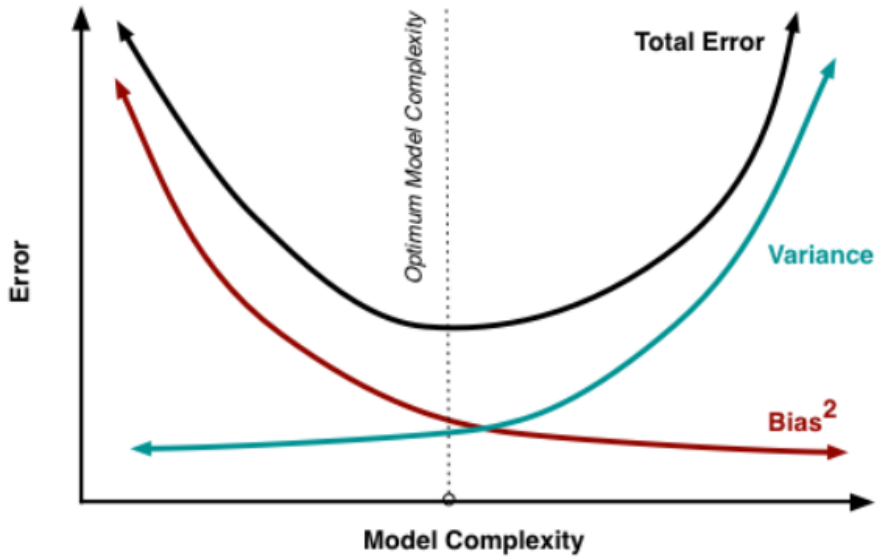
- The Sum of Squares are a representation of the error for our OLS regression model.
- When we discuss linear regression models, prediction errors can be decomposed into two main subcomponents we care about: error due to "bias" and error due to "variance".
- Understanding bias and variance is critical for understanding the behavior of prediction models, but in general what you really care about is overall error, not the specific decomposition.
- Understanding how different sources of error lead to bias and variance helps us improve the data fitting process resulting in more accurate models.

Bias and Variance Tradeoff

- **Error due to Bias:** The error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict.
- **Error due to Variance:** The error due to variance is taken as the variability of a model prediction for a given data point.
- Imagine you can repeat the entire model building process multiple times. The variance is how much the predictions for a given point vary between different realizations of the model.



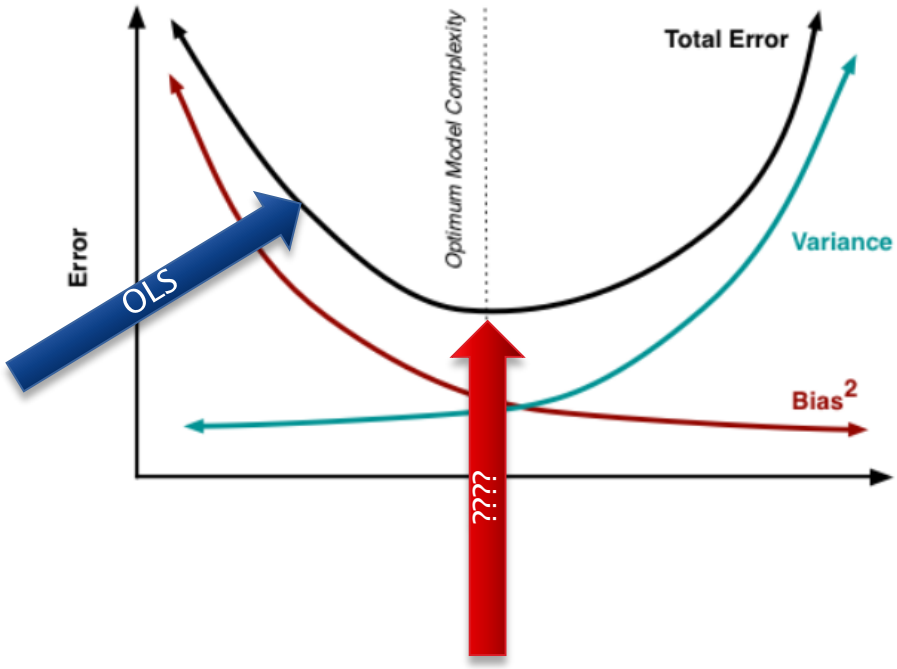
Bias and Variance Tradeoff



- There is a tradeoff between a model's ability to minimize bias and variance. Understanding these two types of error can help us diagnose model results and avoid the mistake of over- or under-fitting.
- The sweet spot for any model is the level of complexity at which the increase in bias is equivalent to the reduction in variance.
- Bias is reduced and variance is increased in relation to model complexity. As more and more parameters are added to a model, the complexity of the model rises and variance becomes our primary concern while bias steadily falls.
- For example, as more polynomial terms are added to a linear regression, the greater the resulting model's complexity will be

Gauss Markov Theorem

- The Gauss Markov theorem states that among all linear unbiased estimates, OLS has the smallest variance.
- This implies that our OLS Estimates have the smallest mean squared error among linear estimators with no bias.
- This begs the question: *Can there be a biased estimator with a smaller mean squared error?*



Shrinkage Estimators

Let's consider something that initially seems crazy:

- We will replace our OLS estimates β_k with something slightly smaller:

$$\beta'_k = \frac{1}{1 + \lambda}\beta_k$$

- If λ is zero, we get the OLS estimates back; if λ gets really large, the parameter estimate approaches a minimal value (zero).

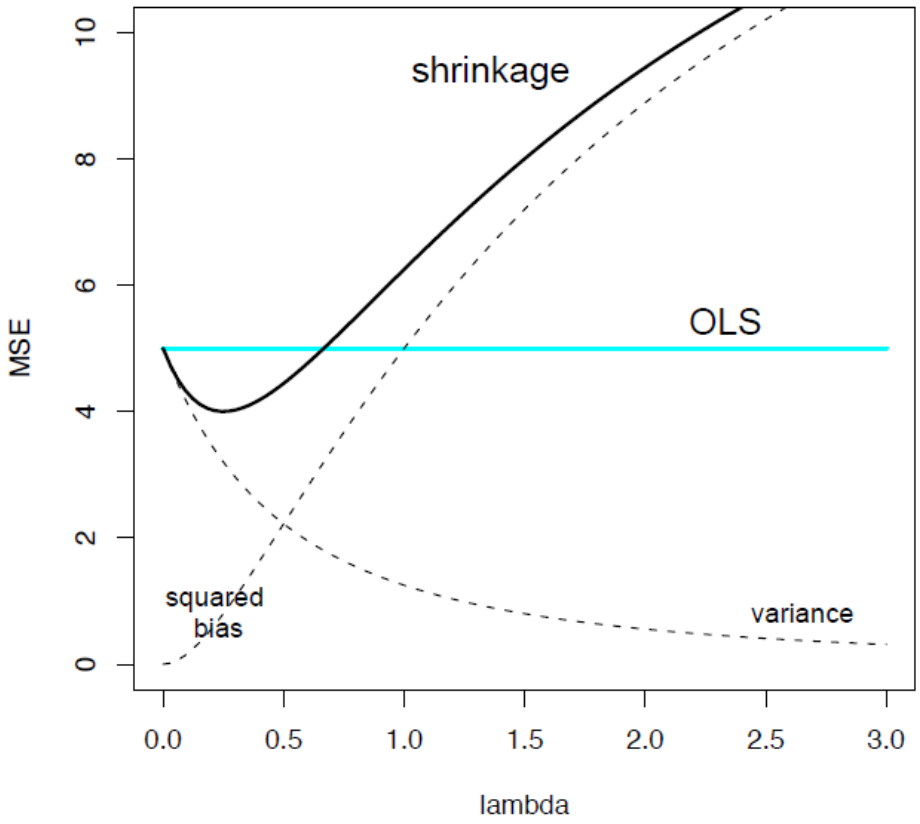
- λ is referred to as the shrinkage estimator (ridge constant).

Shrinkage Estimators

- In principle, with the right choice of λ we can get an estimator with a better MSE.
- The estimate is not unbiased but what we pay for in bias, we make up for in variance.
- To find the minimum λ by balancing the two terms we get the following:

$$\lambda = \frac{p\sigma^2}{\sum {\beta_k}^2}$$

- Therefore, if all of the coefficients are large relative to their variances, we set λ to be small.
- On the other hand, if we have a number of small coefficients, we want to pull them closer to 0 by setting the λ to be large.



Shrinkage Estimators

If we are trying to establish a good value of λ, with the optimal being:

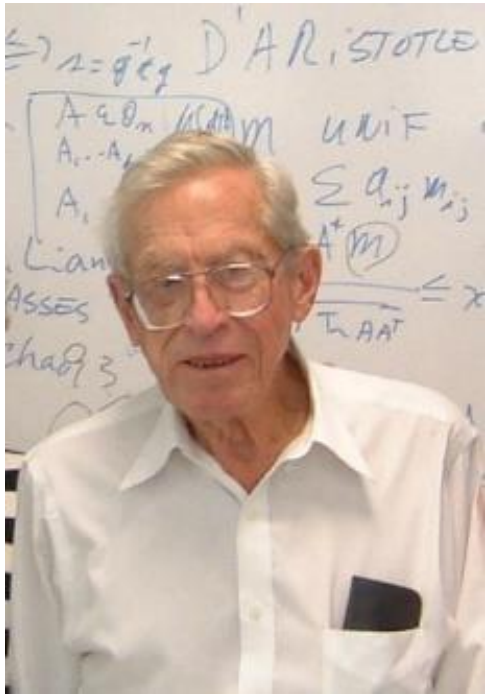
$$\lambda = \frac{p\sigma^2}{\sum {\beta_k}^2}$$

- Do we ever really have access to this information?

Suppose we know σ^2

- In the early 1960's, Charles Stein working with a graduate student Willard James came up with the following specification:

$$\beta'_k = (1 - \frac{(p-2)\sigma^2}{\sum {\beta_k}^2})\beta_k$$



Shrinkage Estimators

- This formula was expanded upon by Stanley Sclove in the late 1960's.
- Stanley's proposal was to shrink the estimates to 0 if we get a negative value

$$\beta'_k = \left(1 - \frac{(p-2)\sigma^2}{\sum {\beta_k}^2}\right)^+ \beta_k$$

- where $(x)^+ = \max(x,0)$
- If σ^2 is unknown, he proposed that we the following criterion:

$$\beta'_k = \left(1 - c\frac{RSS}{\sum {\beta_k}^2}\right)^+ \beta_k$$

- for some value of c

Shrinkage Estimators

- This formula can be re-expressed as the following:

$$F = \frac{\frac{\sum_k \beta_k{}^2}{p}}{\frac{RSS}{(n-p)}}$$

- And then expressing Scolve's estimate as:

$$\beta'_k = \left(1 - \frac{c}{p}\right)^+ \beta_k$$

- This statement reads that we will set the coefficients to 0 unless F > c
- Alternatively, the result shows that we set the coefficients to 0 if we fail an F-test with a significance level set by the value of c.
- If we pass the F-test, then we shrink the coefficients by an amount determined by how strongly the F-statistic protests the null hypothesis.

Shrinkage Estimators

- This preliminary testing procedure acts to either kills coefficients or keeps them (and shrinks them).
- This is kind of like model selection, except that it kills all of the coefficients (unlike the keep or kill rules experience with AIC and BIC).
- We know that simple model selection via AIC and BIC can be applied to regular regressions.
- AIC = n log(RSS) + 2df
- BIC = n log(RSS) + df log(n)
- What about the shrinkage stuff?

Ridge Regression

- Ridge Regression is a modeling technique that works to solve the multi-collinearity problem in OLS models through the incorporation of the shrinkage parameter, λ.
- The assumptions of the model are the same as OLS. Linearity, constant variance, and independence. Normality not need be assumed.
- Additionally, multiple linear regression (OLS) has no manner to identify a smaller subset of important variables.

Ridge Regression

- In OLS regression, the form of the equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$ can be represented in matrix notation as follows:

$$X_t X\beta = XtY$$

- where X is the design matrix having, $[X]_{ij} = x_{ij}$, y is the vector of the response $(y_1,\ldots,y_n)$, and β is the vector of the coefficients $(\beta_1,\ldots, \beta_p)$.

- This equation can be rearranged to show the following:

$$\beta = (X'X)^{-1} X'Y$$

- where R = X'X
- and R is the correlation matrix of independent variables.

Ridge Regression

- Here is the rearranged OLS equation again from the previous slide.

$$\beta = (X'X)^{-1} X'Y$$

- These estimates are unbiased so the expected values of the estimates are the population values. That is,

$$E(\beta') = \beta$$

- The variance-covariance matrix of the estimates is

$$V(\beta') = \sigma^2 R^{-1}$$

- and since we are assuming that the y's are standardized, $\sigma^2 = 1$

Ridge Regression

- Ridge Regression proceeds by adding a small value, λ, to the diagonal elements of the correlation matrix. (This is where ridge regression gets its name since the diagonal of ones may be thought of as a ridge.)

$$\beta = (R + \lambda I)^{-1} X'Y$$

- λ is a positive value less than one (usually less than 0.3).

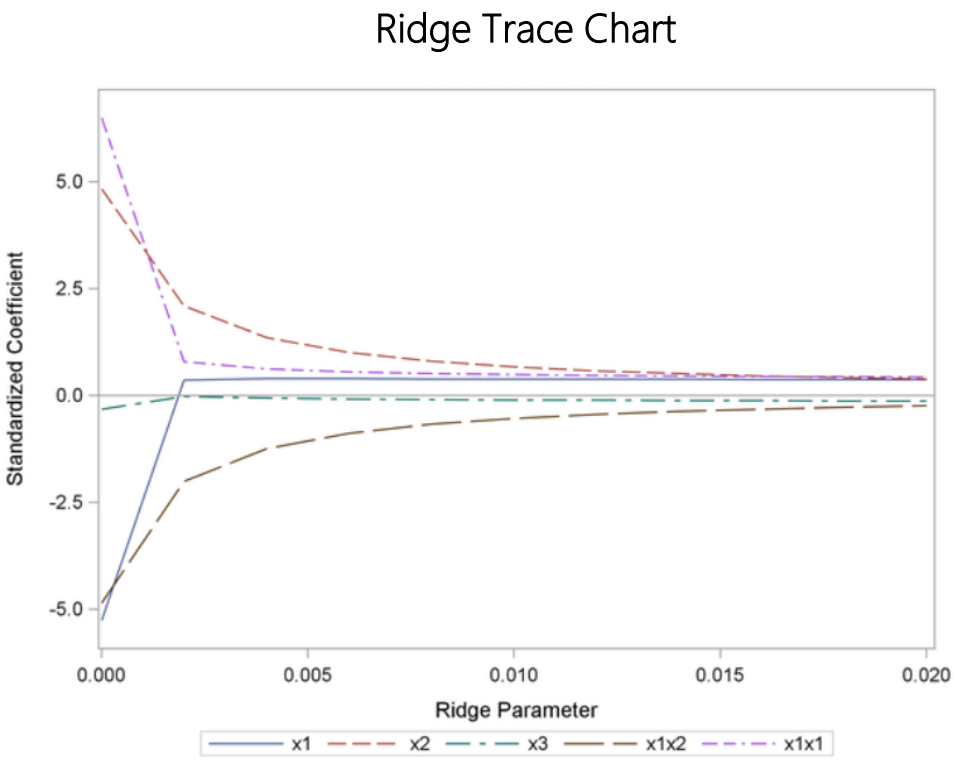
- The amount of bias of the estimator is given by:

$$E(\beta' - \beta) = [(X'X + \lambda I)^{-1} X'X - I]\beta$$

- and the covariance matrix is given by:

$$V(\beta') = (\mathrm{X}'\mathrm{X} + \lambda I)^{-1} X'X (\mathrm{X}'\mathrm{X} + \lambda I)^{-1}$$

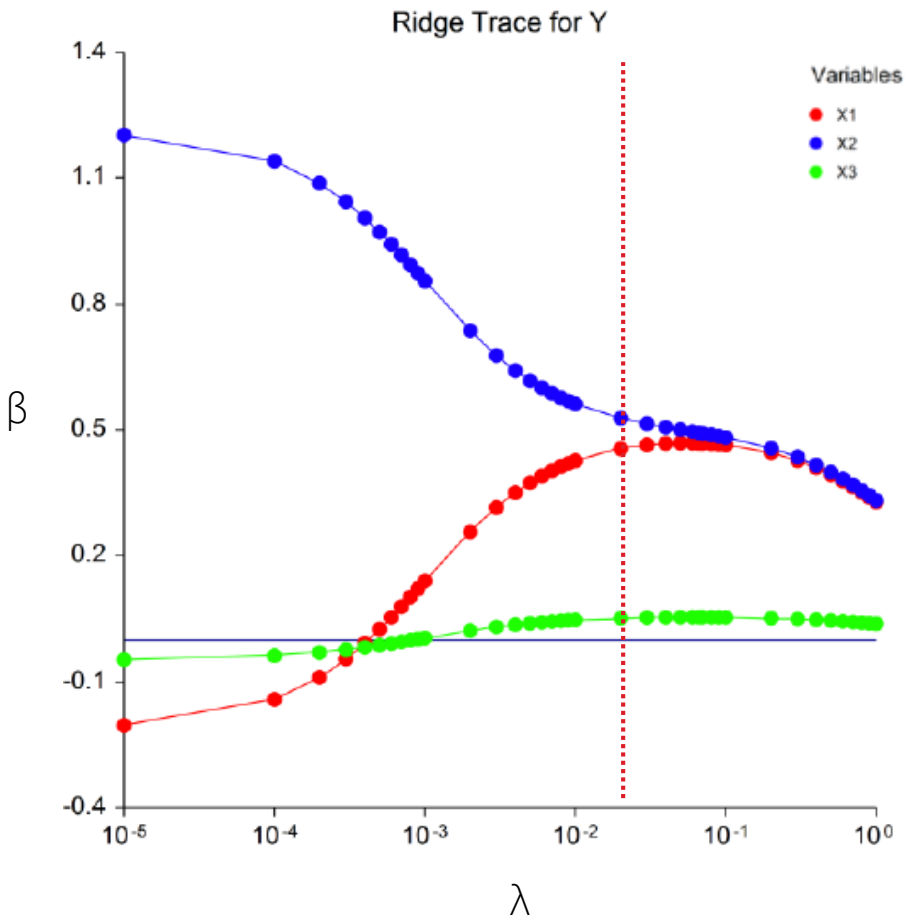
Ridge Trace



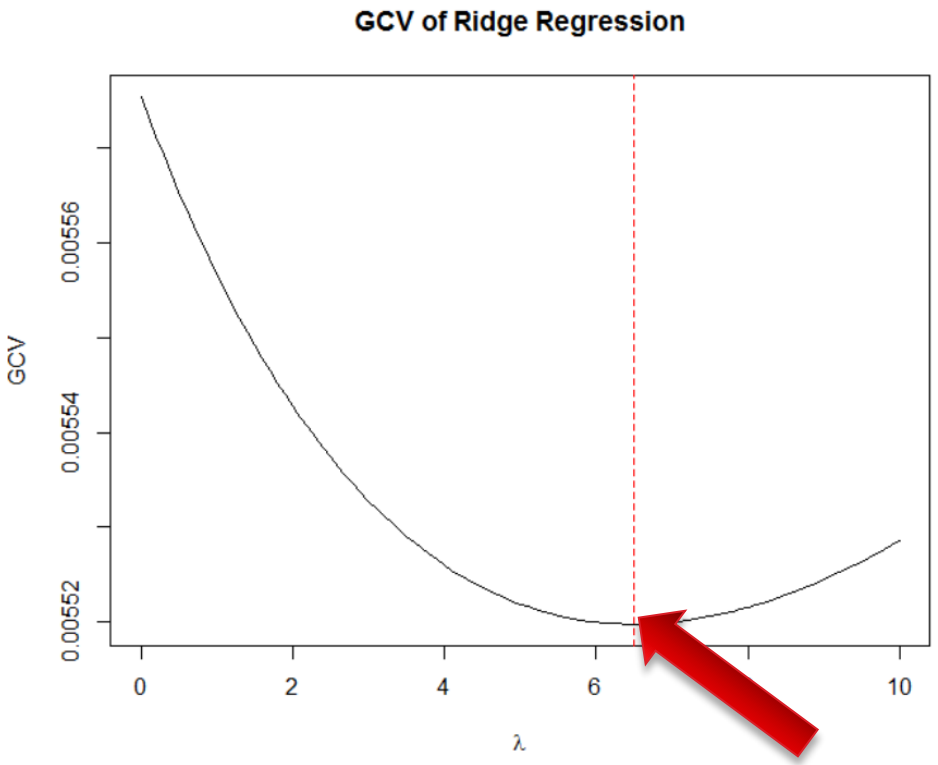
- One of the main obstacles in using ridge regression is choosing an appropriate value of λ. The inventors of ridge regression suggested using a graphic which they called a "ridge trace".
- A ridge trace is a plot that shows the ridge regression coefficients as a function of λ.
- When viewing the ridge trace we are looking for the λ for which the regression coefficients have stabilized. Often the coefficients will vary widely for small values of λ and then stabilize.
- Choose the smallest value of λ possible (which introduces the smallest bias) after which the regression coefficients seem to have remained constant.
- **Note:** Increasing λ will eventually drive the regression coefficients to 0.

Ridge Trace

- In this example, the values of λ are shown on a logarithmic scale. I have drawn a vertical line at the selected value of λ = 0.006.
- We see that λ has little effect until λ is about 0.001 and the action seems to stop somewhere near 0.006.
- Notes: The vertical axis contains points for the least squares solution. These are labeled as 0.000001.
- This was done so that these coefficients can be seen. In fact, the logarithm of zero is -∞ so the least squares values cannot be displayed when the horizontal axis is put in log scale.



Ridge Bias Constant

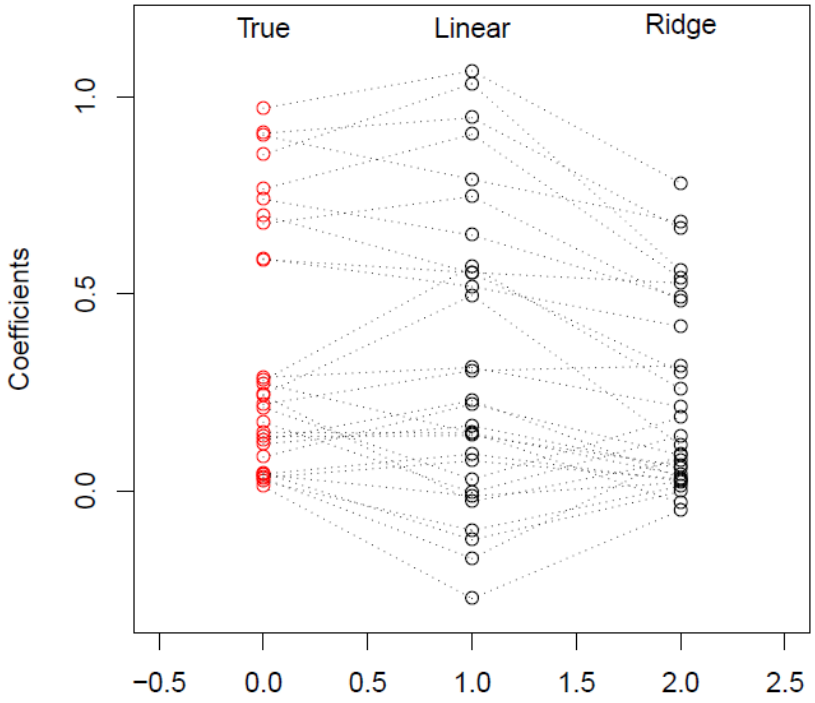


- Alternatively, there is a procedure in R which automatically selects the lowest value for λ.
- The value is calculated using a general cross validation procedure (GVC).
- The example on the right shows the value as 6.5 which is the lowest point on the curve.
- Note: the sequence range for the λ will need to be re-specified for each model that you are building, in case the λ range exceeds your specified bounds.

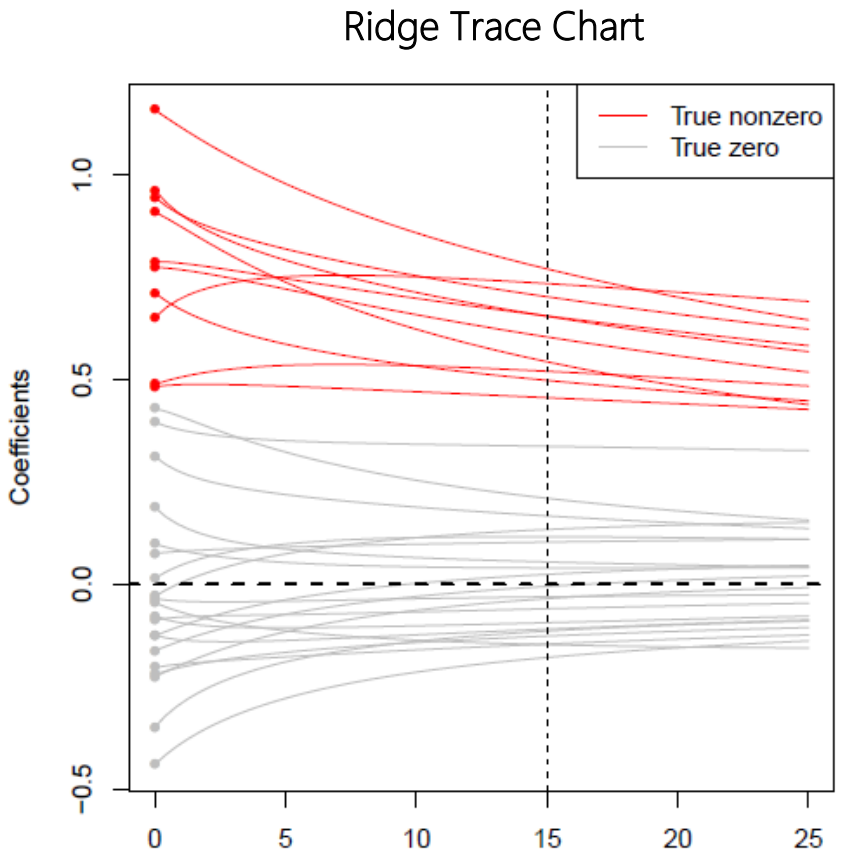
```
# Using R's automatic selection methods to select the biasing constant:
# R calls this constant "lambda"

select(lm.ridge(lpsa~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45,
                data=mydata, lambda = seq(0,1,0.001)))
```

Scale in Ridge Regression

- Here is a visual representation of the ridge coefficients for λ versus a linear regression.
- We can see that the size of the coefficients (penalized) has decreased through our shrinking function, ℓ_2 (more on this later).
- It is also important to point out that in ridge regression we usually leave the intercept un-penalized because it is not in the same scale as the other predictors.
- The λ is unfair if the predictor variables are not on the same scale.
- Therefore, if we know that the variables are not measured in the same units, we typically center and scale all of the variables before building a ridge regression.



Variable Selection



- The problem of picking out the relevant variables from a larger set is called variable selection.
- Suppose there is a subset of coefficients that are identically zero. This means that the men response doesn't depend on these predictors at all.
- The red paths on the plot are the true non zero coefficients, the grey paths are the true zeros.
- The vertical dash line is the point which ridge regression's MSE starts losing to linear regression.
- Note: the grey coefficient paths are not exactly zero; they are shrunken, but non zero.

Variable Selection

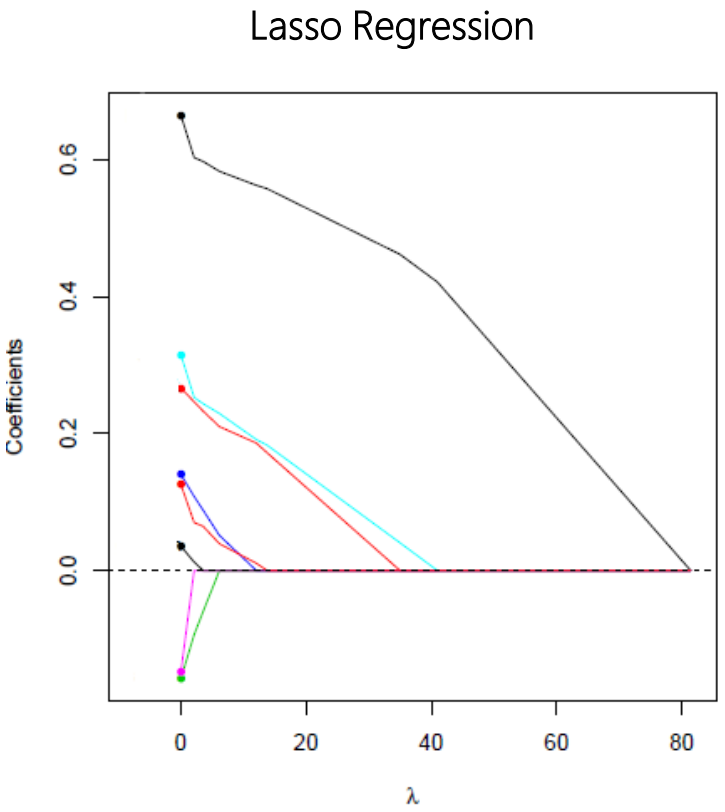
- We can show that ridge regression doesn't set the coefficients exactly to zero unless $\lambda = \infty$, in which case they are all zero.
- Therefore, ridge regression cannot perform variable selection.
- Ridge regression performs well when there is a subset of true coefficients that are small or zero.
- It doesn't do well when all of the true coefficients are moderately large, however, will still perform better than OLS regression.

LASSO

- The lasso combines some of the shrinking advantages of ridge regression with variable selection.
- Lasso is an acronym for “Least Absolute Selection and Shrinkage Operator”.
- The lasso is very competitive with the ridge regression in regards to prediction error.
- The only difference between the lasso and ridge regression is that the ridge ℓ_2 uses the $\| \beta \|_2^2$ penalty where the lasso ℓ_1 uses the $\| \beta \|_1$ penalty (more later).
- Even though these ℓ_1 and ℓ_2 look similar, their solutions behave very differently.

LASSO

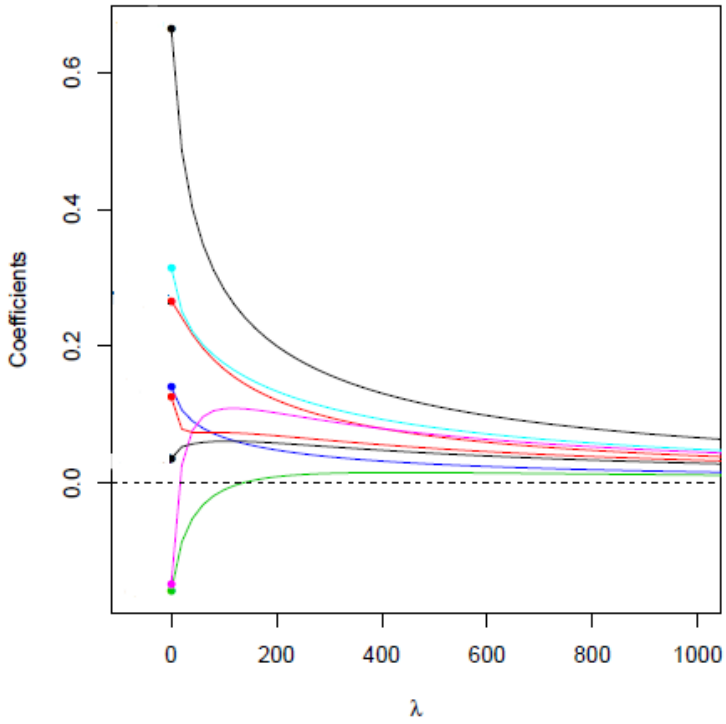
- The tuning parameter λ controls the strength of the penalty and like ridge regression we get the β'^{lasso} = the linear regression estimate when λ = 0, and β'^{lasso} when λ = ∞.
- For λ in between these two extremes, we are balancing 2 ideas: fitting a linear model of y on X, and shrinking the coefficients.
- The nature of the penalty ℓ_1 causes some of the coefficients to be shrunken to zero exactly.
- This is what makes lasso different than ridge regression. It is able to perform variable selection in the linear model.
- Important: As λ increases, more coefficients are set to zero (less variables selected), and among non-zero coefficients, more shrinkage is employed.



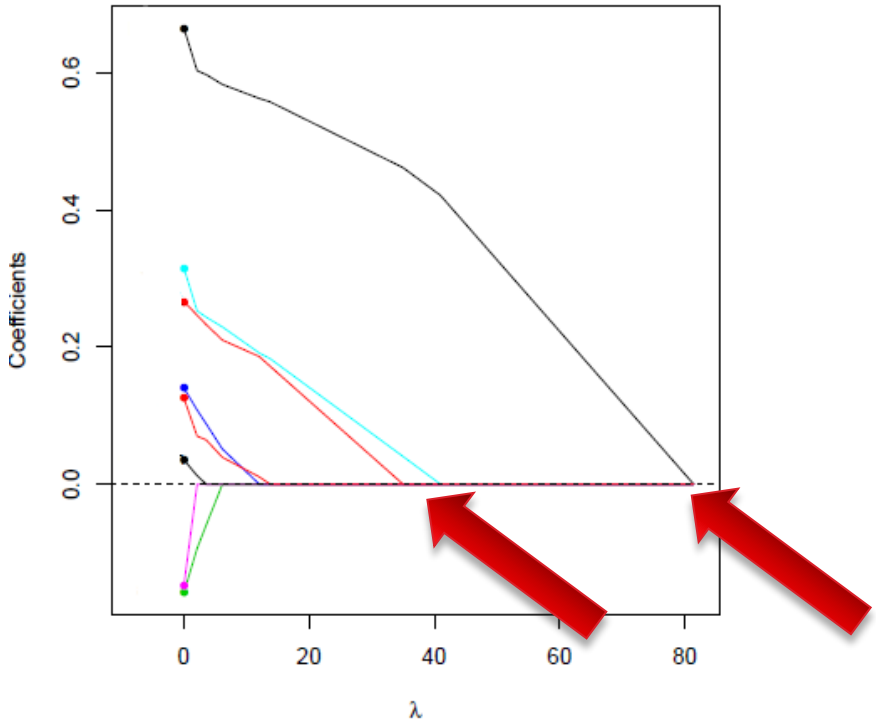
LASSO

- Because the lasso sets the coefficients to exactly zero it performs variable selection in the linear model.

Ridge Regression



Lasso Regression

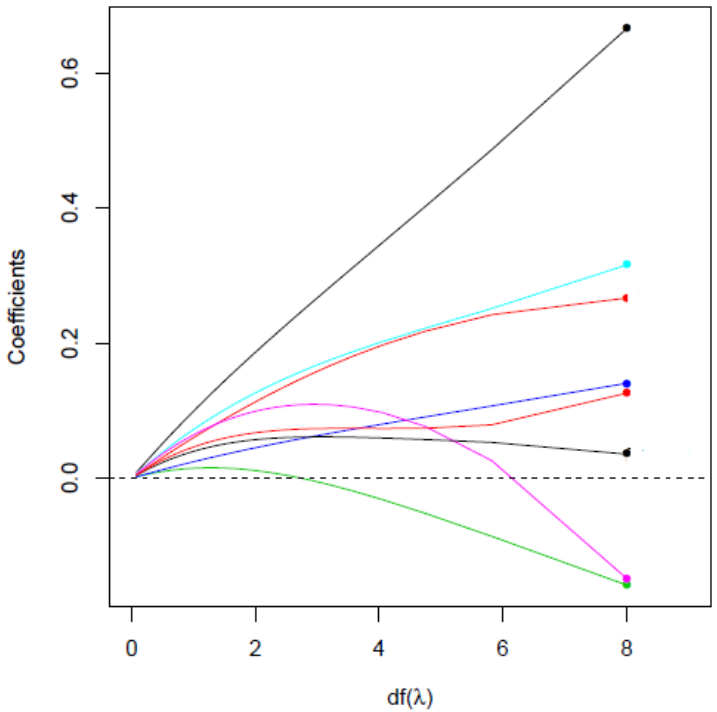


The variables with the largest λ values in LASSO that converge to 0 indicate the most desirable variables for the model.

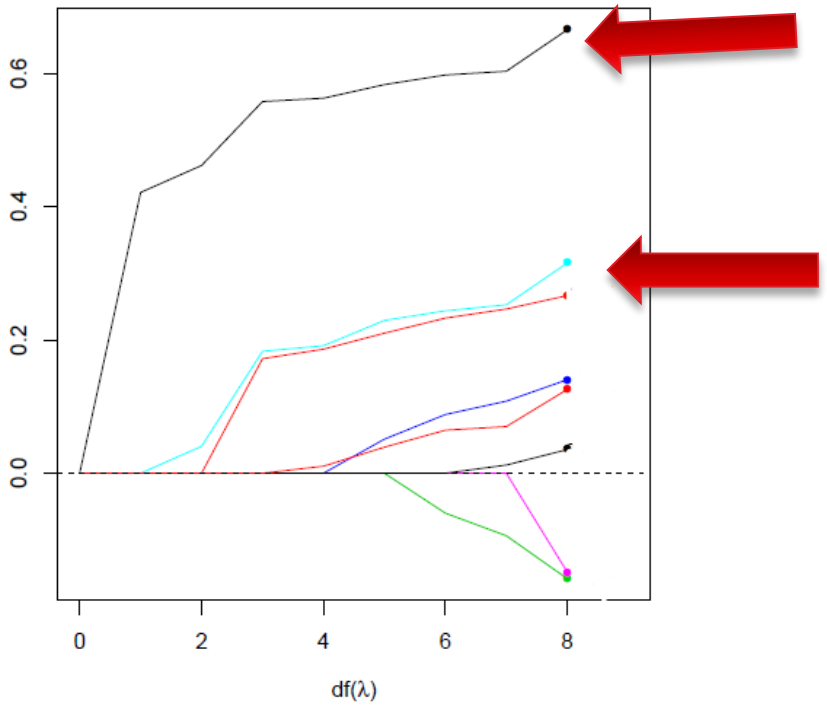
LASSO

- We can also use plots of the degrees of freedom (df) to put different estimates on equal footing.

Ridge Regression



Lasso Regression



Constrained Form

- It can be helpful to think about our penalty ℓ_1 and ℓ_2 parameters in the following form:

$$\hat{\beta}^{\text{lasso}} = \text{argmin } \|y - X\beta|_2^2 + \lambda \|\beta\|_1 \quad \leftarrow \ell_1$$

$$\hat{\beta}^{\text{ridge}} = \text{argmin } \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \quad \leftarrow \ell_2$$

- We can think of this formula now in a constrained (penalized) form:

$$\hat{\beta}^{\text{lasso}} = \text{argmin } \|y - X\beta|_2^2 \quad \text{subject to } \|\beta\|_1 \leq t$$

$$\hat{\beta}^{\text{ridge}} = \text{argmin } \|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_2^2 \leq t$$

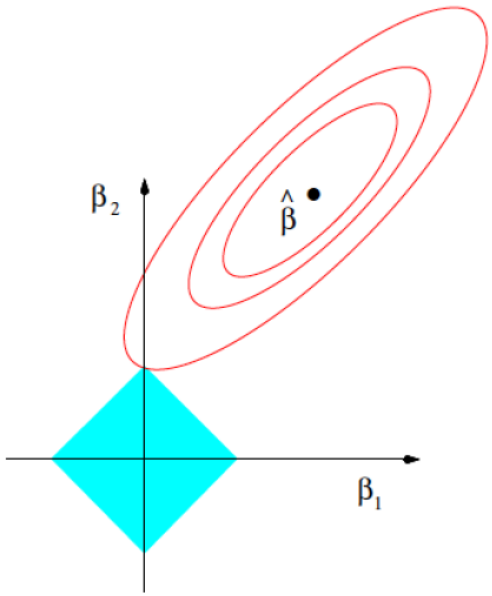
- t is a tuning parameter (which we have been calling λ earlier)

- The usual OLS regression solves the unconstrained least squares problem; these estimates constrain the coefficient vector to lie in some geometric shape centered around the origin.

Constrained Form

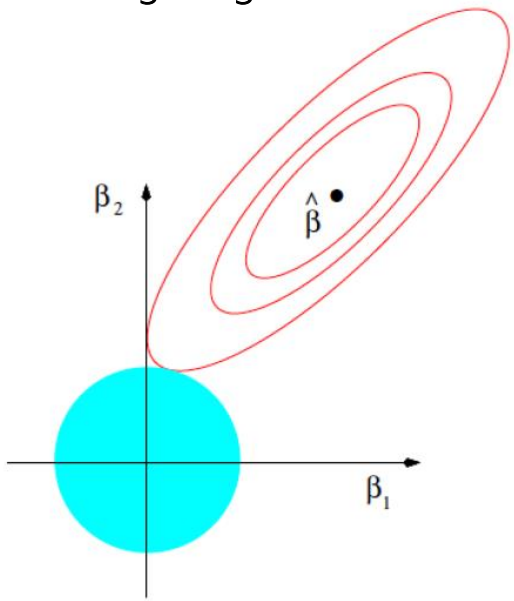
- This generally reduces the variance because it keeps the estimate close to zero. But the shape that we choose really matters!!!

Lasso Regression



The contour lines are the least squares error function. The blue diamond is the constraint region for the lasso regression.

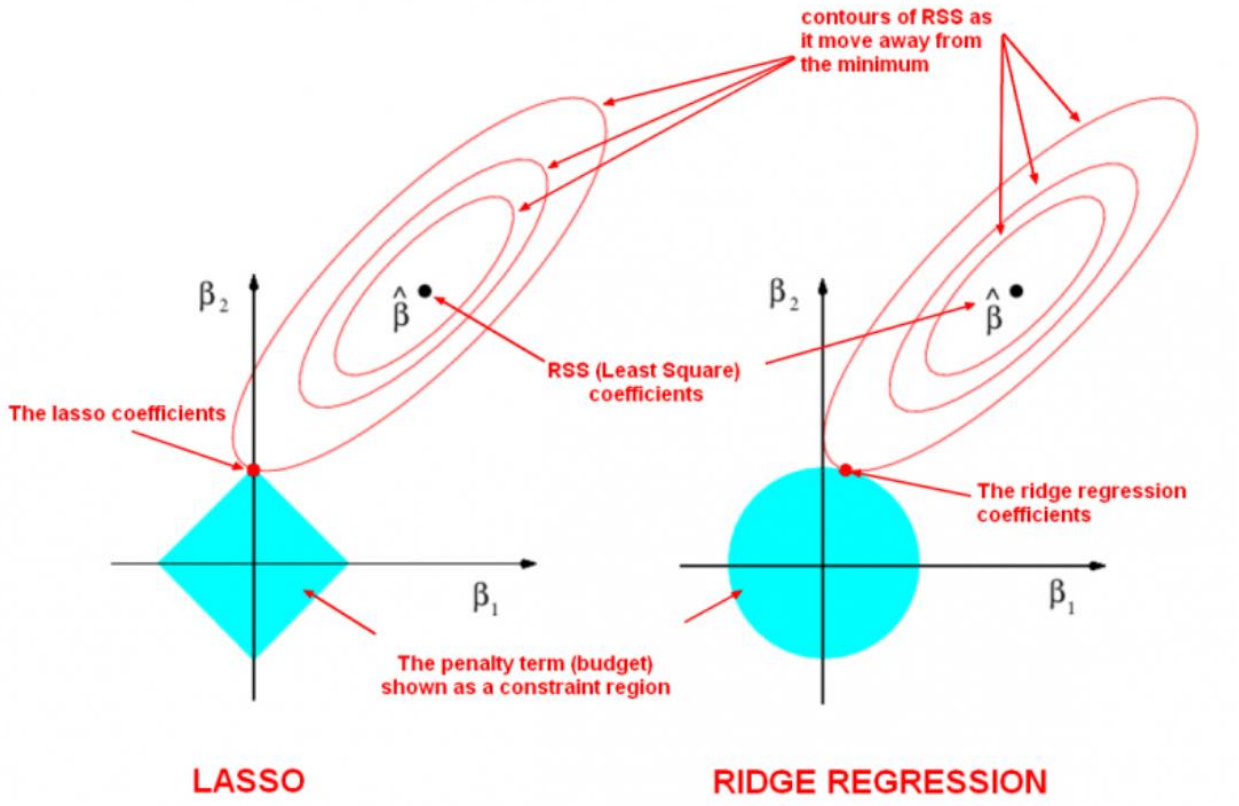
Ridge Regression



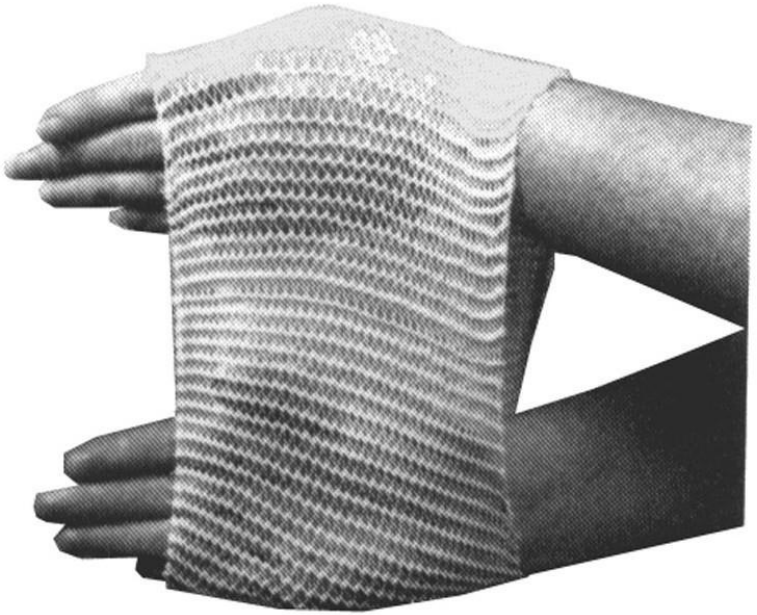
The contour lines are the least squares error function. The blue circle is the constraint region for the lasso regression.

Constrained Form

- Here is a more detailed breakdown:



Elastic Net



- When we are working with high dimensional data (datasets with a large number of independent variables), correlations between the variables can be high resulting in multicollinearity.
- These correlated variables can sometimes forms groups or clusters of correlated variables.
- There are many times where we would want to include the entire group in the model selection if one variable has been selected.
- This can be thought of as an elastic net catching a school of fish instead of singling out a single fish.
- An example of data would be the Leukemia dataset which contains 7129 genes, with correlation, and a tumor type.

Elastic Net

- The total number of variables that the lasso variable selection procedure is bound by the total number of samples in the dataset.
- Additionally, the lasso fails to perform grouped selection. It tends to select one variable from a group and ignore the others.
- The elastic net forms a hybrid of the ℓ_1 and ℓ_2 penalties:

$$\hat{\beta}^{\text{ridge}} = \text{argmin } \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$$\hat{\beta}^{\text{lasso}} = \text{argmin } \|y - X\beta|_2^2 + \lambda \|\beta\|_1$$

$$\hat{\beta}^{\text{elastic}} = \text{argmin } \|y - X\beta\|_2^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

Elastic Net

- Ridge, Lasso, and Elastic Net are all part of the same family with the penalty term of:

$$P_\alpha = \sum_{i=1}^{p} [\frac{1}{2}(1-\alpha)b_j^2 + \alpha|b_j|]$$

PENALTY BOX

- If the α = 0 then we have a Ridge Regression
- If the α = 1 then we have the LASSO
- If the 0 < α < 1 then we have the elastic net

Elastic Net

$$P_\alpha = \sum_{i=1}^{p} [\frac{1}{2}(1-\alpha)b_j^2 + \alpha|b_j|]$$

- The specification of the elastic net penalty above is actually considered a naïve elastic net.
- Unfortunately, the naïve elastic net does not perform well in practice.
- The parameters are penalized twice with the same α level. (this is why it is called naïve)
- To correct this we can use the following:

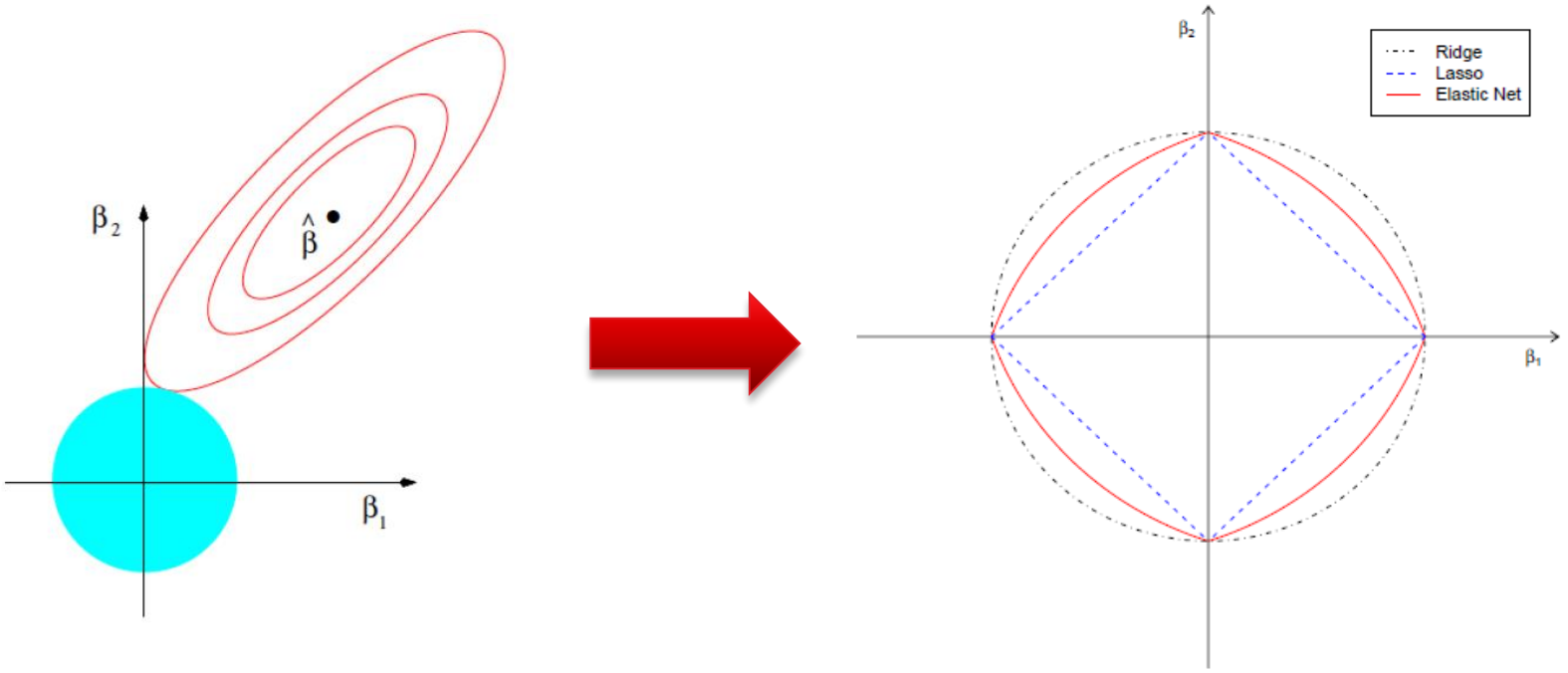
$$\text{Penalty} = (1-\alpha)|\boldsymbol{\beta}|_1 + \alpha|\boldsymbol{\beta}|^2$$

$$= \lambda_2|\boldsymbol{\beta}|^2 + \lambda_1|\boldsymbol{\beta}|_1 \quad \text{where} \quad \alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$$

$$\hat{\boldsymbol{\beta}}(\text{elastic net}) = (1+\lambda_2)\hat{\boldsymbol{\beta}}(\text{naive elastic net})$$

Elastic Net - Constraint

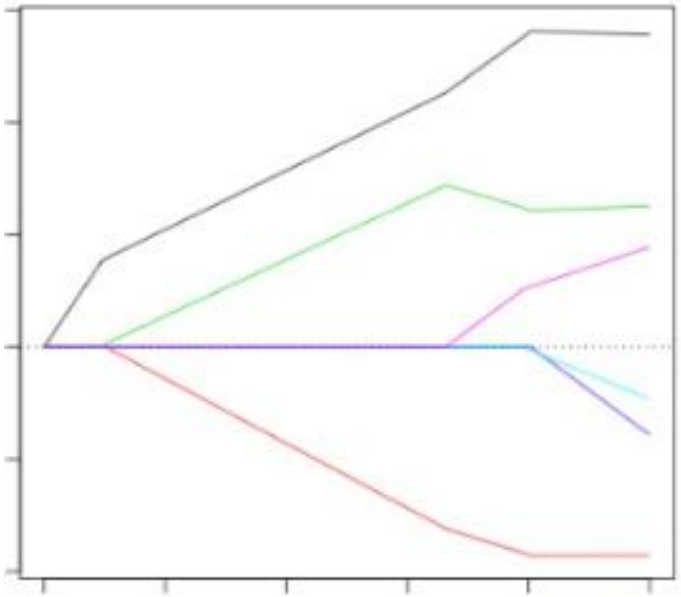
- Here is a visualization of the constrained region for the elastic net.



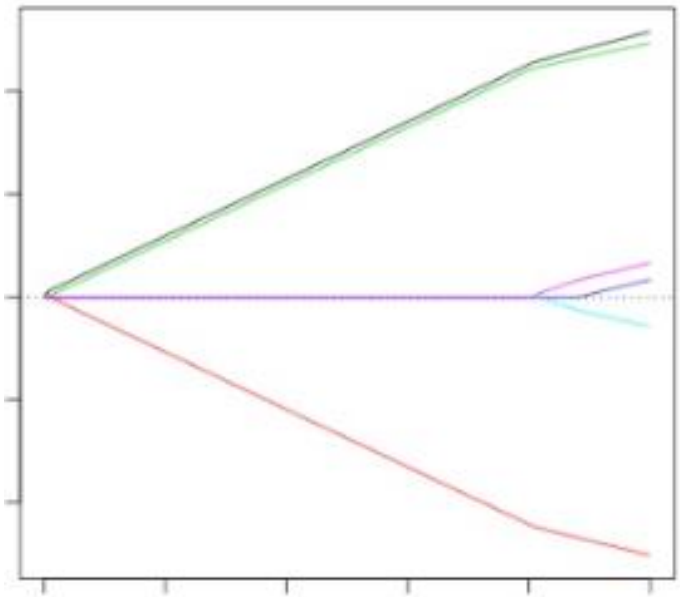
Elastic Net

- We can see that the elastic net organizes the coefficients (lasso rope) into organized groups forming the “framework” for the elastic net.

Lasso



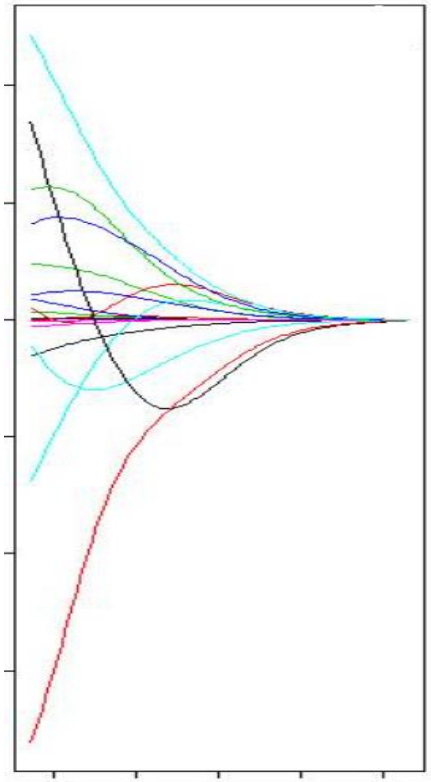
Elastic Net



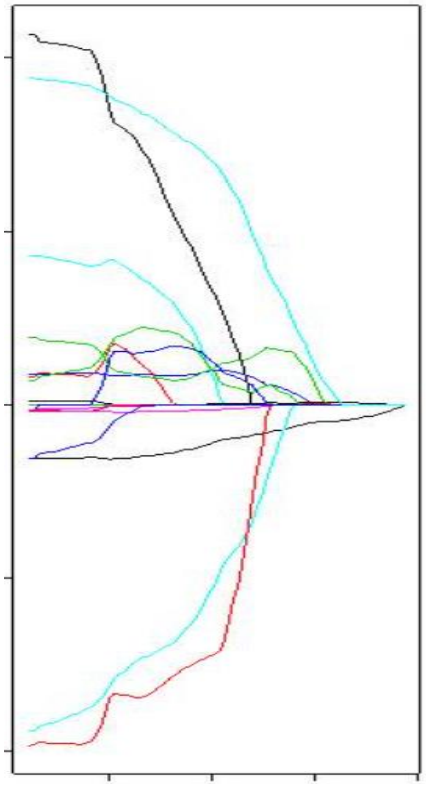
Ridge, Lasso, & Elastic Nets

- Lets take a final look at visualizations for the Ridge Regression, Lasso, and Elastic Nets.

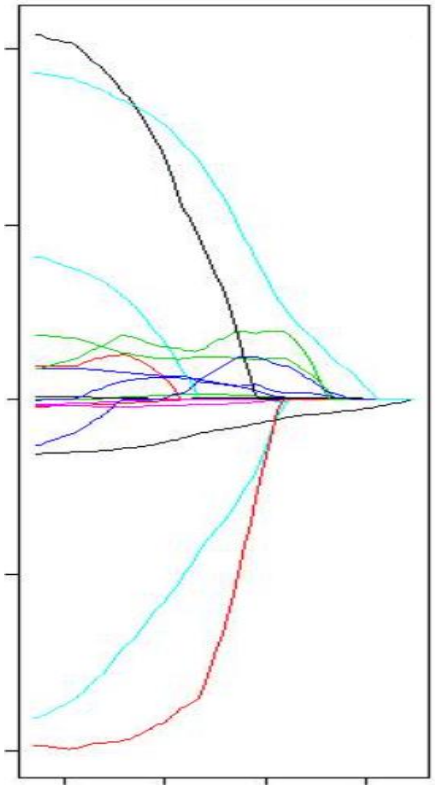
Ridge



Lasso



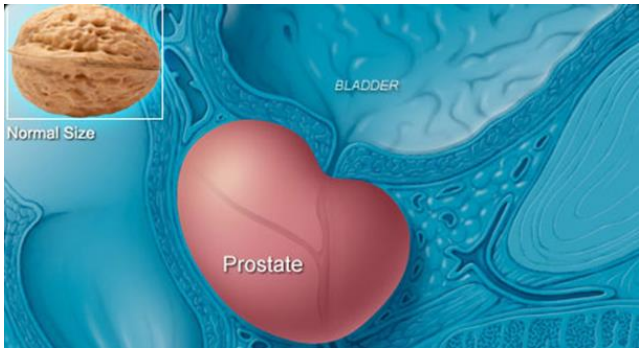
Elastic Net



Modern Regression Example: Prostate Cancer

Understanding the Data

- There are many forms of cancer which humans contract throughout their lifetime and prostate cancer is a prevalent form in males.
- By having an understanding of the variables and measurements which impact the development of the cancer, this will aid researchers in developing medical treatment options.
- The dataset we will be working with contains various measurements between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy.
- The goal for this exercise will be to examine various predictive models that can be leveraged to help model the relationships.



Understanding the Data

- Here is a description of the variables within the dataset:

Variable	Description
lcavol	the log of the cancer volume
lweight	the log of the prostates weight
age	age of the patient
lbph	the log of the benign prostatic hyperplasia amount
svi	seminal vesicle invasion
lcp	the log of the capsular penetration
gleason	Gleason score
pgg45	percentage Gleason scores 4 or 5
lpsa	the log of the prostate specific antigen

- Our goal is to develop various regression models (ridge, lasso, and elastic net) for the lpsa based upon the other variables.

Understanding the Data

- First, lets take a look at the raw data in the table.

lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
-0.57982	2.769459	50	-1.38629	0	-1.38629	6	0	-0.43078
-0.99425	3.319626	58	-1.38629	0	-1.38629	6	0	-0.16252
-0.51083	2.691243	74	-1.38629	0	-1.38629	7	20	-0.16252
-1.20397	3.282789	58	-1.38629	0	-1.38629	6	0	-0.16252
0.751416	3.432373	62	-1.38629	0	-1.38629	6	0	0.371564
-1.04982	3.228826	50	-1.38629	0	-1.38629	6	0	0.765468
0.737164	3.473518	64	0.615186	0	-1.38629	6	0	0.765468
0.693147	3.539509	58	1.536867	0	-1.38629	6	0	0.854415
-0.77653	3.539509	47	-1.38629	0	-1.38629	6	0	1.047319
0.223144	3.244544	63	-1.38629	0	-1.38629	6	0	1.047319
0.254642	3.604138	65	-1.38629	0	-1.38629	6	0	1.266948
-1.34707	3.598681	63	1.266948	0	-1.38629	6	0	1.266948
1.61343	3.022861	63	-1.38629	0	-0.59784	7	30	1.266948
1.477049	2.998229	67	-1.38629	0	-1.38629	7	5	1.348073

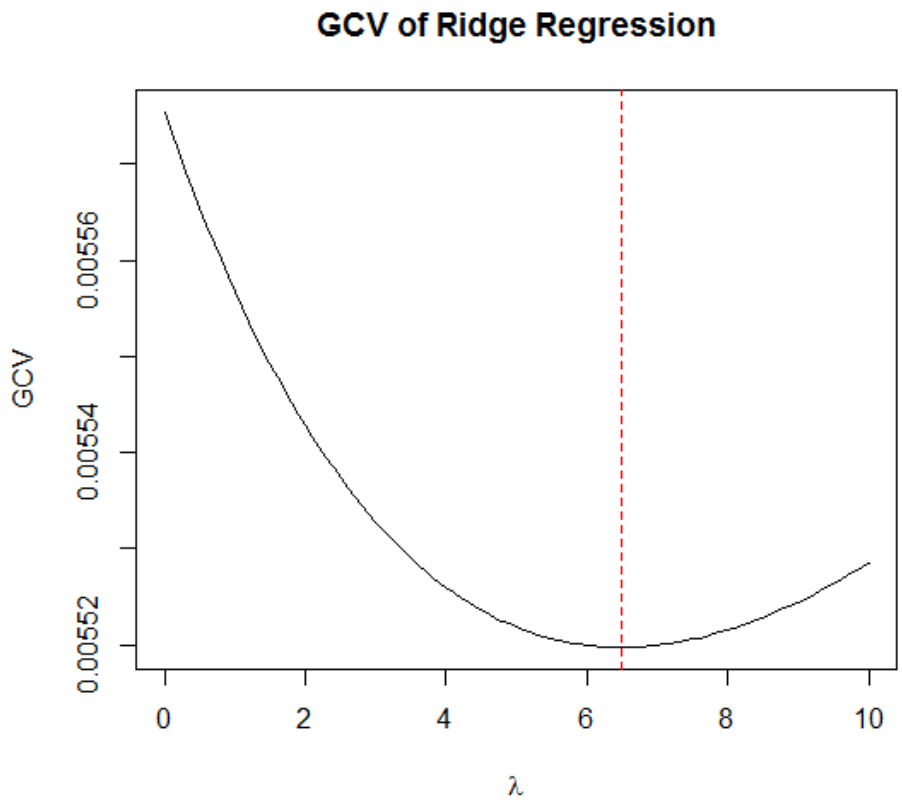
Linear Regression Model Selection

- Lets first create a linear regression model to assess the predictive performance.
- We will start off by building a model with all of the variables included and then pare down the model.
- The variables within the model was simplified by removing those variables with a p > 0.05, as indicated in the graphic.
- In order to assess the predictive performance, we calculated the Mean Square Error (MSE) for this linear regression model which was 0.49262.

Coefficients:				
Parameter	**Estimate**	**Std. Error**	**t value**	**Pr>|t|**
Intercept	0.669399	1.296381	0.516	6.07E-01
lcavol	0.587023	0.08792	6.677	2.11E-09
lweight	0.454461	0.170012	2.673	0.00896
age	-0.01964	0.011173	-1.758	0.08229
lbph	0.107054	0.058449	1.832	0.0704
svi	0.766156	0.244309	3.136	0.00233
lcp	-0.10547	0.091013	-1.159	0.24964
gleason	0.045136	0.157464	0.287	0.77506
pgg45	0.004525	0.004421	1.024	0.30885

Coefficients:				
Parameter	**Estimate**	**Std. Error**	**t value**	**Pr>|t|**
Intercept	-0.26807	0.5435	-0.493	6.23E-01
lcavol	0.55164	0.07467	7.388	6.30E-11
lweight	0.50854	0.15017	3.386	0.00104
svi	0.66616	0.20978	3.176	0.00203

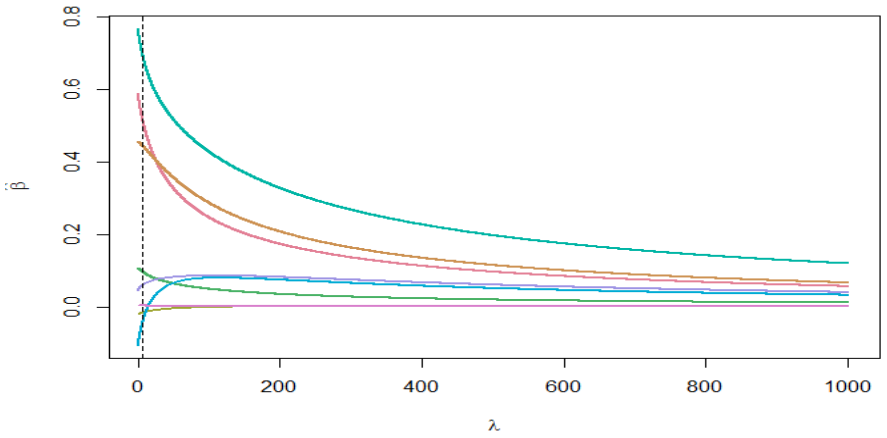
Ridge Regression Model

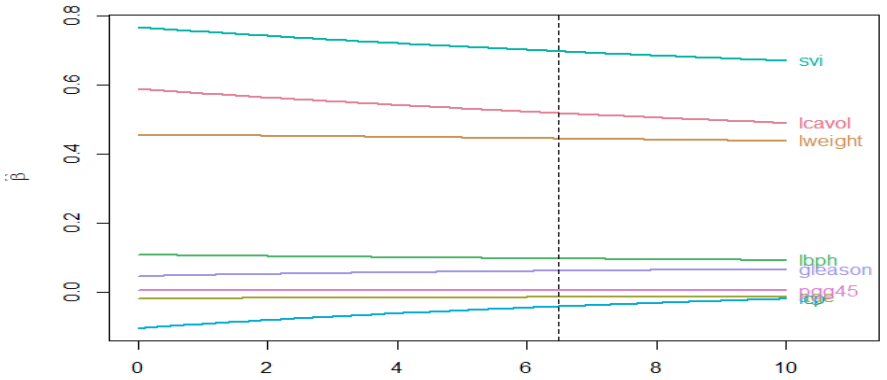


- Now lets see if we can further improve upon the MSE by exploring the utilization of a ridge regression model.
- The first step we will utilize will be to calculate the minimum value of λ to use in the ridge model.
- After flexing the boundaries of the λ values, we have determined that the λ of 6.5 would be the ideal value.

Ridge Regression Model

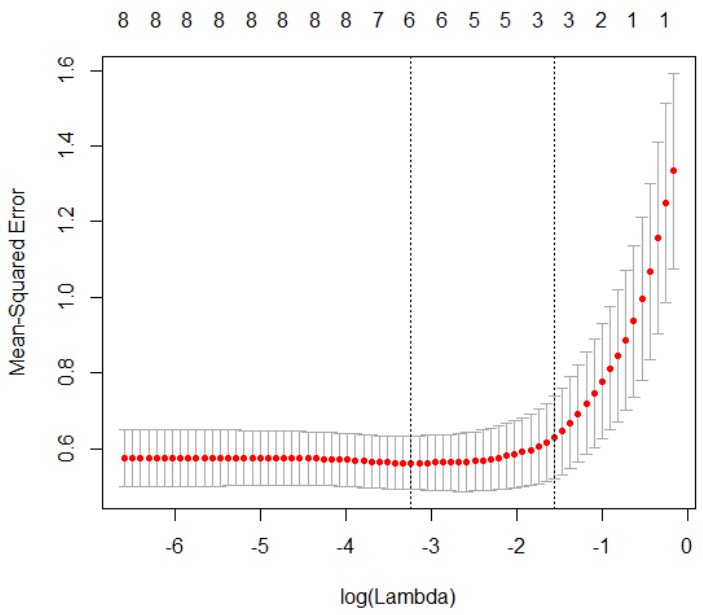
- We should then create a ridge trace to see the convergence of the coefficients to zero. This ridge trace is showing the λ with a range of 0 – 1000.
- The next ridge trace shows a subset of the original ridge trace but isolating the values of λ between 0 – 10.
- The ideal value of λ = 6.5 is indicated by the vertical dashed line.
- The MSE of the model was calculated and determined to be 0.4601.

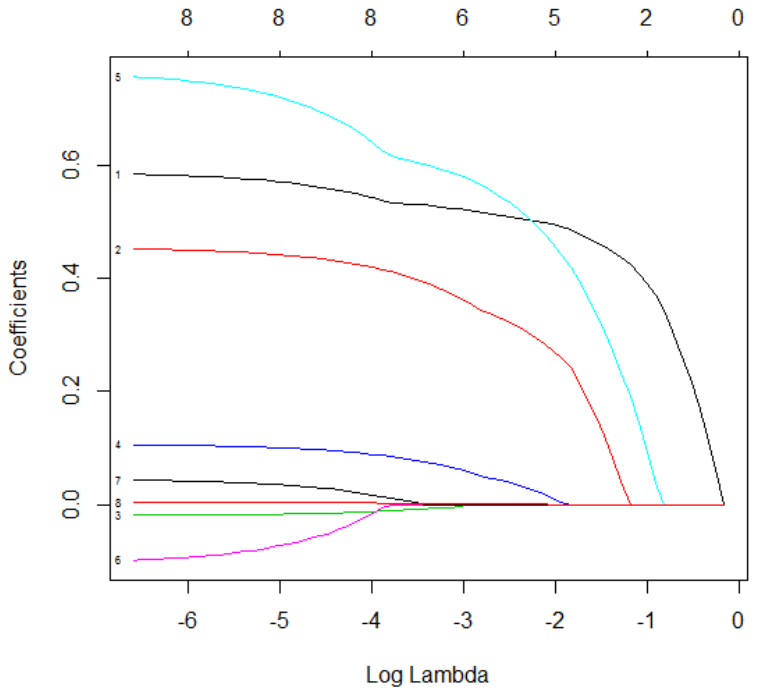




LASSO Model

- Perhaps we can further improve the ridge regression MSE through incorporating the LASSO?
- We utilized the R package glmnet for the LASSO by establishing the alpha parameter =1.



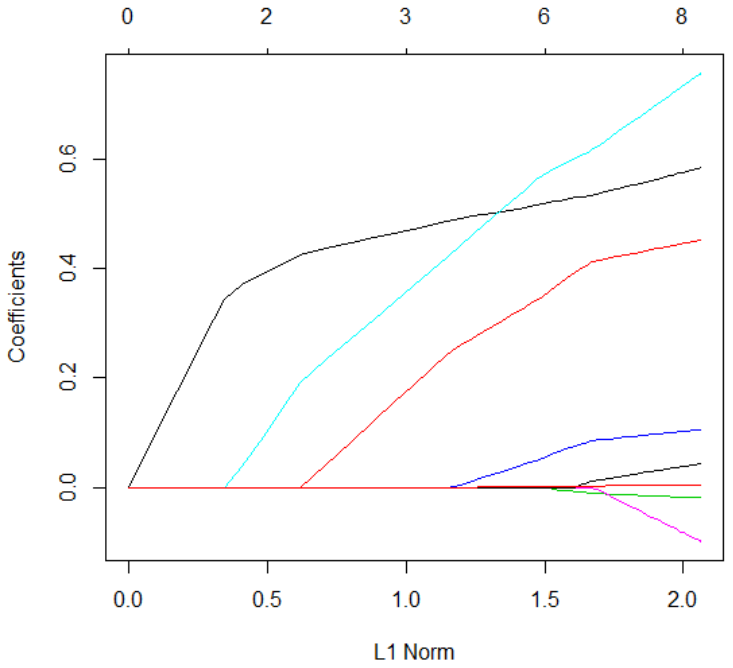


Minimum λ = 0.039

LASSO Model

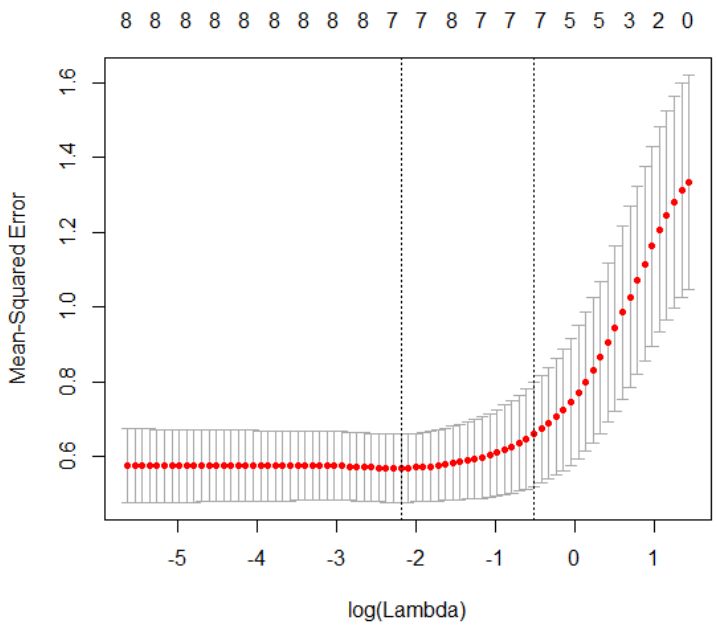
- A comparison of the shrinking of the coefficients between the ridge regression and LASSO can give us some clues about the significance of variables.
- The variables lcp and gleason had been reduced to 0, which effectively eliminates them from the LASSO model.
- Also, please note that the variables age and pgg45 are very close to a 0 value which indicates that they are less significant than the other variables.
- The regression model utilizing the LASSO had produced a MSE of 0.4725.

Variable	Ridge	Lasso
Intercept	0.489196	0.564073
lcavol	0.517031	0.525988
lweight	0.443541	0.380376
age	-0.01553	-0.00585
lbph	0.096061	0.069318
svi	0.696091	0.592326
lcp	-0.04255	0
gleason	0.060571	0
pgg45	0.00352	0.002184

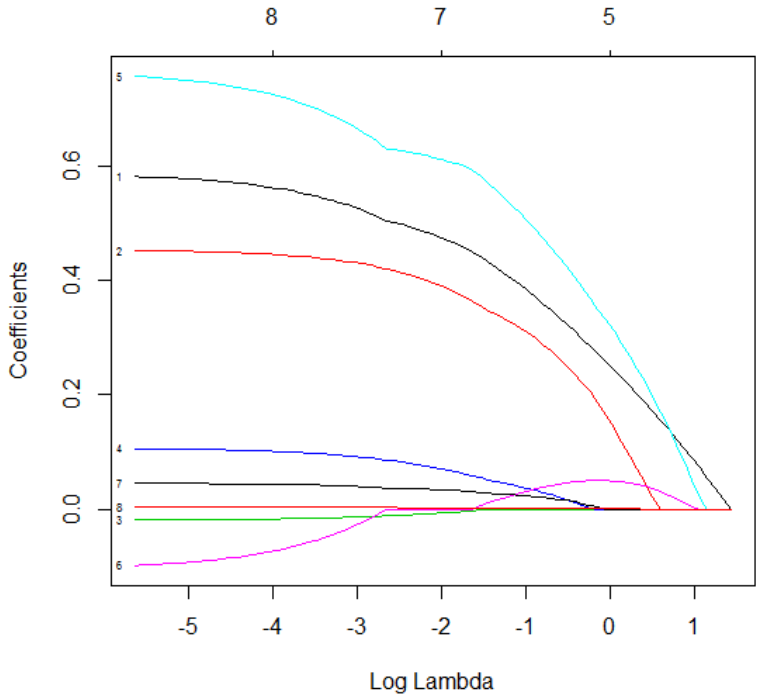


Elastic Net Model

- Finally, lets assess a regression model by utilizing an elastic net.
- We built the elastic net model using the R package glmnet by establishing the alpha parameter = 0.2.

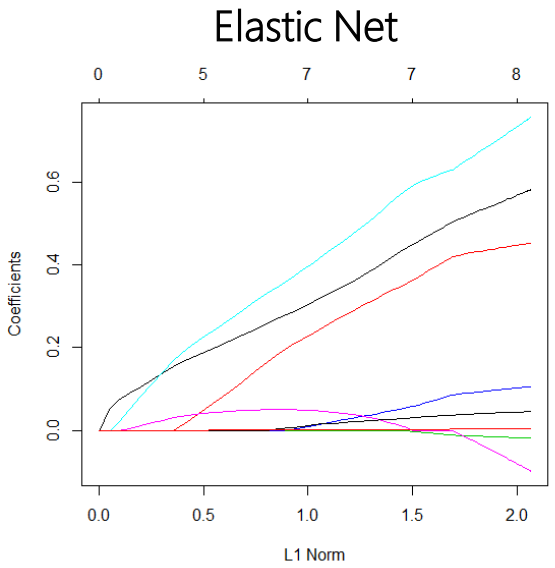


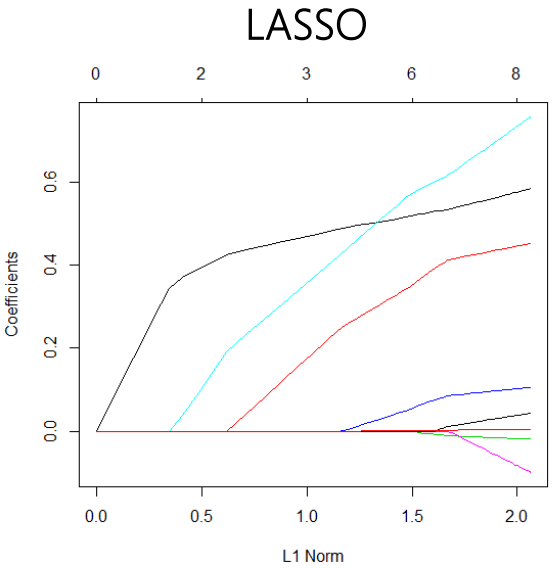
Minimum λ = 0.112



Elastic Net Model

- We first established the minimum lambda to be 0.112 for our elastic net.
- Notice that the rigidness of the shape of coefficients of the elastic net as compared to the LASSO when comparing the penalized L1 norm.
- This is a nice visual way to highlight the differences between the two techniques for non technical audiences.
- The regression model utilizing the elastic net had produced a MSE of 0.47115.





Model Comparisons

- Here is a comparison of the new coefficients values and the predictive performance rank between the regression methods.
- The coefficient values highlighted in red are equal to 0 and have been removed from the final model.

Model Coefficients β

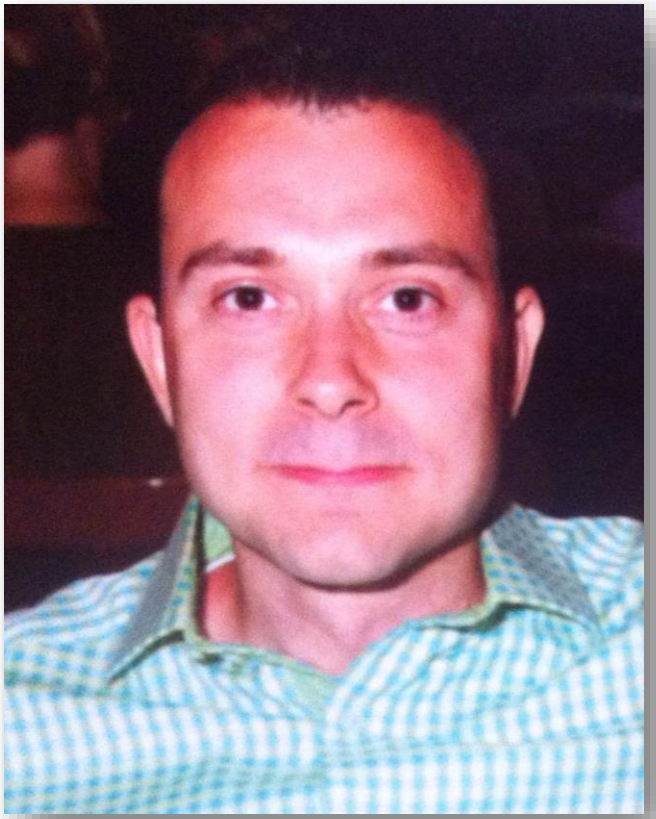
Variables	Ridge	Lasso	Elastic Net
Intercept	0.489	0.564	0.454
lcavol	0.517	0.526	0.485
lweight	0.444	0.380	0.401
age	-0.016	-0.006	-0.008
lbph	0.096	0.069	0.076
svi	0.696	0.592	0.619
lcp	-0.043	0.000	0.000
gleason	0.061	0.000	0.035
pgg45	0.004	0.002	0.003

Predictive Performance (MSE)

Model	MSE	Rank
Ridge Regression	0.4601	1
Elastic Net	0.4716	2
Lasso	0.4725	3
Reduced Linear	0.4926	4

About Me

- Reside in Wayne, Illinois
- Active Semi-Professional Classical Musician (Bassoon).
- Married my wife on 10/10/10 and been together for 10 years.
- Pet Yorkshire Terrier / Toy Poodle named Brunzie.
- Pet Maine Coons' named Maximus Power and Nemesis Gul du Cat.
- Enjoy Cooking, Hiking, Cycling, Kayaking, and Astronomy.
- Self proclaimed Data Nerd and Technology Lover.



Acknowledgements

- http://www.stat.ucla.edu/~cocteau/stat120b/lectures/
- http://scott.fortmann-roe.com/docs/BiasVariance.html
- http://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf
- http://web.stanford.edu/~hastie/TALKS/enet_talk.pdf
- http://www.biecek.pl/WZUR2009/AgnieszkaProchenka2009.pdf
- http://www.slideshare.net/ShangxuanZhang/ridge-regression-lasso-and-elastic-net
- http://www.stat.cmu.edu/~ryantibs/datamining/lectures/16-modr1.pdf
- http://www.stat.cmu.edu/~ryantibs/datamining/lectures/17-modr2.pdf
- http://www.slideshare.net/ShangxuanZhang/ridge-regression-lasso-and-elastic-net
- http://www4.stat.ncsu.edu/~post/reading/josh/LASSO_Ridge_Elastic_Net_-_Examples.html
- http://www.moseslab.csb.utoronto.ca/alan/glmnet_presentation.pdf

Fine