Multiple Linear Regression - Data Analysis and Visualization

Dr. Farshid Alizadeh-Shabdiz

Spring 2021

Multiple Linear Regression (MLR)

The equation for the simple linear regression line is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

- y is the response or dependent variable
- $x_1, x_2, \ldots, x_k$ are the explanatory or independent variables
- β_0 is the intercept (the value of y when the $x_1, x_2, \ldots, x_k$ are set to 0)
- β_1 is the slope (the expected change in y for each one-unit change in x_1 after adjusting for $x_2, \ldots, x_k$)
- β_k is the slope (the expected change in y for each one-unit change in x_k after adjusting for $x_1, x_2, \ldots, x_{k-1}$)
- e is the random error which we assume is normally distributed with **a mean of 0 and a variance of σ^2**

Regression Equation

The equation for the multiple linear regression line is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

In the least-squares regression, the estimates are selected in such a way that the following quantity is **minimized**:

$$(y - \hat{y})^2 = \left[y - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k\right)\right]^2$$

Assessing the Fit of the Regression Line - coefficient of determination

The coefficient of determination represents the proportion (percentage) of the variation in the response variable explained by the multiple regression model.

- Coefficient of determination (R^2) is the same as SLR - in MLR we have more variables.
- Given that there are more than one independent variable in this setting, the **coefficient of determination is not simply the squared correlation coefficient.**

In MLR, R^2 is referred to as the multiple R-squared.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\text{Reg SS}}{\text{Total SS}}$$

Inference Global F test

- We use the ANOVA table and are interested in testing the **alternative hypothesis that at least one of the slope parameters is different than 0.**
- If this test confirms that there is at least one slope parameter that is different than 0, then **subsequent F-tests or t-tests** can be used to assess whether each individual slope parameter is different than 0.
- In MLR, the F-test for the model is referred to as the **global test.**

Inference Global F test

- We use the ANOVA table and are interested in testing the **alternative hypothesis that at least one of the slope parameters is different than 0.**
- If this test confirms that there is at least one slope parameter that is different than 0, then **subsequent F-tests or t-tests** can be used to assess whether each individual slope parameter is different than 0.
- In MLR, the F-test for the model is referred to as the **global test.**

Difference to SLR is that, here $k > 1$. The exact value of k depends on the number of variables in the model.

	SS (Sum of Squares)	df (degrees of freedom)	MS (Mean Square)	F-statistic	p-value
Regression	Reg SS	Reg df$= k$	Reg MS = Reg SS/Reg df	F=Reg MS/Res MS	$P(F_{\text{Reg df,Res df},\alpha} > F)$
Residual	Res SS	Res df$= n-k-1$	Res MS = Res SS/Res df		
Total	Total SS = Reg SS + Res SS				

Inference - Anova Table Components

	SS (Sum of Squares)	df (degrees of freedom)	MS (Mean Square)	F-statistic	p-value
Regression	Reg SS	Reg df$= k$	Reg MS = Reg SS/Reg df	F=Reg MS/Res MS	$P(F_{\text{Reg df},\text{Res df},\alpha} > F)$
Residual	Res SS	Res df$= n-k-1$	Res MS = Res SS/Res df		
Total	Total SS = Reg SS + Res SS				

- Reg SS$=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$, Res SS$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$, Total SS$=\sum_{i=1}^{n}(y_i-\bar{y})^2$
- **Reg *df* = *k* equals to the number of predictors in the model.**
- **Res *df* = *n* – *k* – 1** equals to sample size minus the number of predictors in the model minus 1.
- Reg MS = Reg SS/Reg df (the regression mean square)
- Res MS = Res SS/Res df (the residual mean square)
- F = Reg MS/Res MS
- p-value = the probability that the observed value of test statistic or a more extreme value could have been observed by chance

F-Test for MLR

$$F = \frac{\text{MS Reg}}{\text{MS Res}}$$

F-distribution with k and $n-k-1$ degrees of freedom under H_0.

The decision rule for a level α test is:

Reject $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ if $F \geq F_{k,n-k-1,\alpha}$

Otherwise, do not reject H_0

where **$F_{k,n-k-1,\alpha}$ is the value from the F-distribution** with

- k degree of freedom (numerator) and
- $n-k-1$ degrees of freedom (denominator) and
- associated with a right-hand tail probability of α.

MLR Inference t-test

If the overall model is significant, then the significance could be attributed to any one of the independent variables.

Perform testing on each individual parameter to identify the relative contribution of each independent variable.

In order to test each if $\beta_i = 0$ after controlling for the other independent variables in the model, we use a t statistic:

$$t = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}}$$

where

$SE_{\hat{\beta}_i}$ the standard error of the estimate of (in the regression model with the other independent variables included) which follows a t-distribution with n-k-1 degrees of freedom under H_0.

MLR Inference t-test

The decision rule for a two-sided level α test is:

- **Reject $H_0 : \beta_i = 0$ if $|t| \geq t_{n-k-1,\alpha/2}$**
- Otherwise, do not reject $H_0 : \beta_i = 0$

where

$$t_{n-k-1,\alpha/2}$$

is the value from the t-distribution with $n - k - 1$ degrees of freedom and associated with a right hand tail probability of $\alpha/2$.

MLR Inference t-test Confidence Interval

We can calculate the two-sided $100\%(1-\alpha)$ confidence interval for using the following formula:

$$\hat{\beta}_i \pm t_{n-k-1,\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_i}$$

MLR Inference t-test Confidence Interval

We can calculate the two-sided $100\%(1-\alpha)$ confidence interval for using the following formula:

$$\hat{\beta}_i \pm t_{n-k-1,\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_i}$$

We can say with $100\% \times (1-\alpha)$ confidence that the true value of is between

$$\hat{\beta}_i - t_{n-k-1,\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_i} \text{ and } \hat{\beta}_i + t_{n-k-1,\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_i}$$

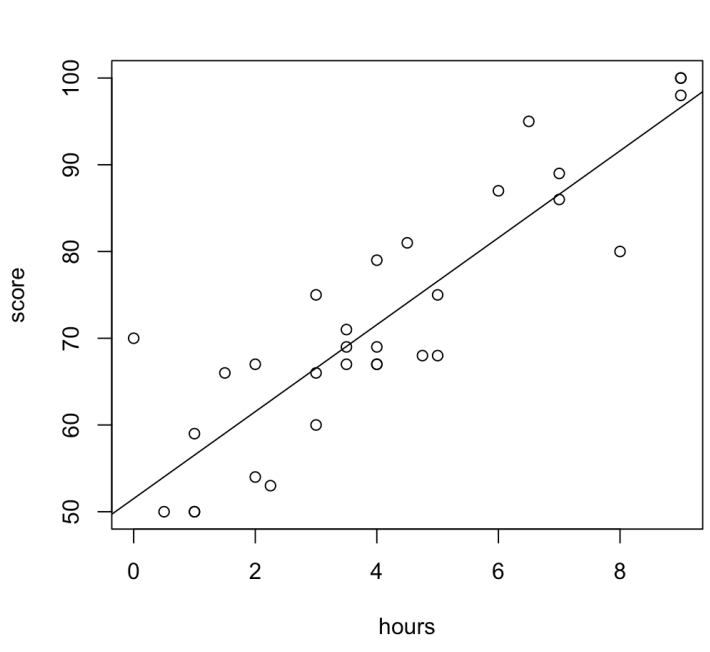
after controlling for the other independent variables in the model.

Overfitting

Example of Overfitting

Data: number of hours studied and score of 31 students.

Linear regression estimate



	name	hours	score
1	Allen	8.0	80
2	Brown	2.0	67
3	Cole	9.0	98
4	Collins	9.0	100
5	Cooper	7.0	86
6	Cox	6.5	95

Learning From the Entire Data

- Perfect training with zero error

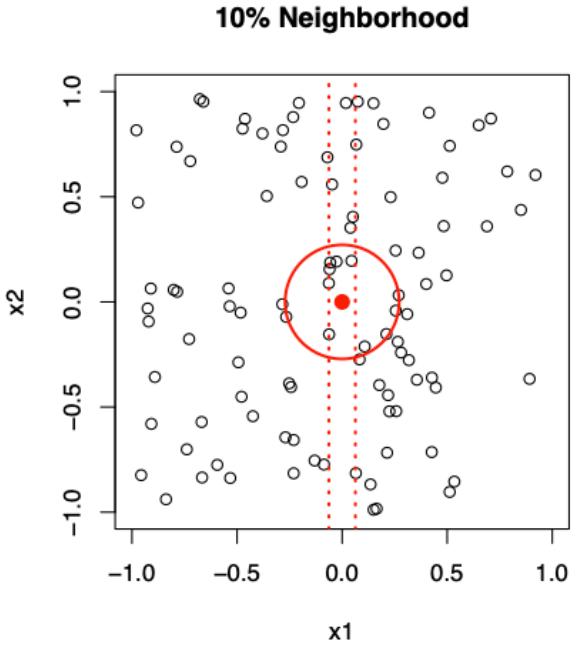
  $R^2 = 1!$ PERFECT!

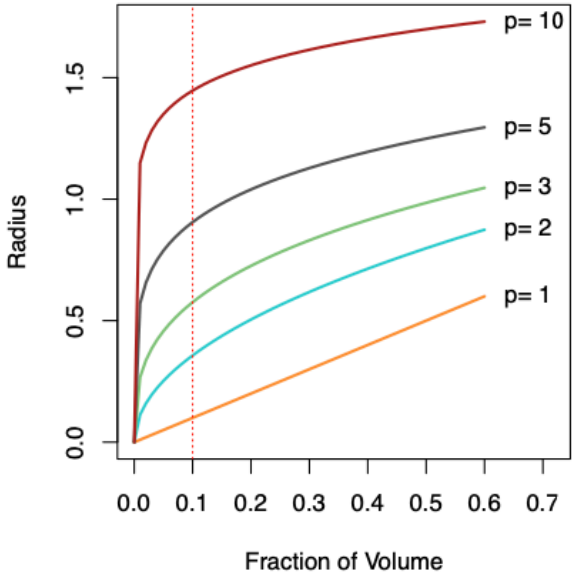
```
Residuals:
ALL 31 residuals are 0: no residual degrees of freedom!

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)        80         NA      NA       NA
nameBrown         -13         NA      NA       NA
nameCole           18         NA      NA       NA
nameCollins        20         NA      NA       NA
nameCooper          6         NA      NA       NA
nameCox            15         NA      NA       NA
nameHall          -13         NA      NA       NA
nameHans          -21         NA      NA       NA
nameHoward          9         NA      NA       NA
nameJeffers       -10         NA      NA       NA
nameJohnson       -30         NA      NA       NA
nameJones         -14         NA      NA       NA
nameKing            1         NA      NA       NA
nameKnight        -11         NA      NA       NA
nameLee            -9         NA      NA       NA
nameMartin        -14         NA      NA       NA
nameMiller         -5         NA      NA       NA
nameMoore         -20         NA      NA       NA
nameMorris        -13         NA      NA       NA
nameMurphy          7         NA      NA       NA
nameReed           -5         NA      NA       NA
nameSmith         -30         NA      NA       NA
nameStewart       -12         NA      NA       NA
nameTaylor        -27         NA      NA       NA
nameThomas        -26         NA      NA       NA
nameThompson      -11         NA      NA       NA
nameWalker        -13         NA      NA       NA
nameWard           20         NA      NA       NA
nameWilliams      -30         NA      NA       NA
nameWright        -12         NA      NA       NA
nameYoung          -1         NA      NA       NA
hours              NA         NA      NA       NA
id                 NA         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
```

The curse of dimensionality





Trade offs

- Prediction accuracy versus interpretability
- Parsimony/Occam's razon vs black box
- Simple model involving less variable is more preferred compare to a black-box model involving all the possible parameters
  - Linear regression models are easy to interpret but some other complex methods are not

How to Assess Multi Parameter Model

Choosing the Optimum Model

- The best model gives us the lowest error over the population and not only training set,
- So R^2 or Residual error is not the best parameter
  - Might results to overfitting!

Adjusted R^2

R Squared is the estimate of the variability of the response variable y given a particular value of the explanatory variable x. It is a statistic and depends on two parameters, 1. number of samples, 2. number of variables in the model (in SLR =1).

We want to be more conservative and adjust (reduce) it to state claims that are more true.

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right)$$

- n is the number of points in your data sample.
- k is the number of independent regressors, i.e., the number of variables in the model (in SLR k=1)
- Always $R^2_{adj} \leq R^2$
- In SLR for large n, $R^2_{adj} \approx R^2$

Another way of writing R^2

- Adjusted R2 can be written as follows which make it easier to understand

$$R^2_{adj} = 1 - \left(\frac{Res\ SS / (n-k-1)}{Total\ SS / (n-1)}\right)$$

Note: simple equation and no need to calculate standard deviation of noise and also works for $n<p$.

Assessing a Model

- Mallow's Cp

$$C_p = \frac{1}{n}(ResSS + 2k\sigma^2)$$

In which

a. k: number of parameters used in the model
b. σ^2: is standard deviation of total error for the full model
c. ResSS: residual error sum square of the small model
d. n: number of samples

- BIC (or Bayesian Information Criterion)

$$BIC = \frac{1}{n}(ResSS + \ln(n)k\sigma^2)$$

In which *n* is # of observations, and *k* is # of parameters.
Note: for $n > 7$ we have $ln(n) > 2$.
Therefore, BIC penalizes more variables more.

A Concern – Training vs Test Data

- The approach minimizes MSE on a training set, e.g. with linear regression we choose the line with minimum MSE.

- But the goal is how well the model works on a new data "**Test Data**".

- There is no guarantee that the smallest training MSE will have the smallest test data MSE.

Assessing a Model

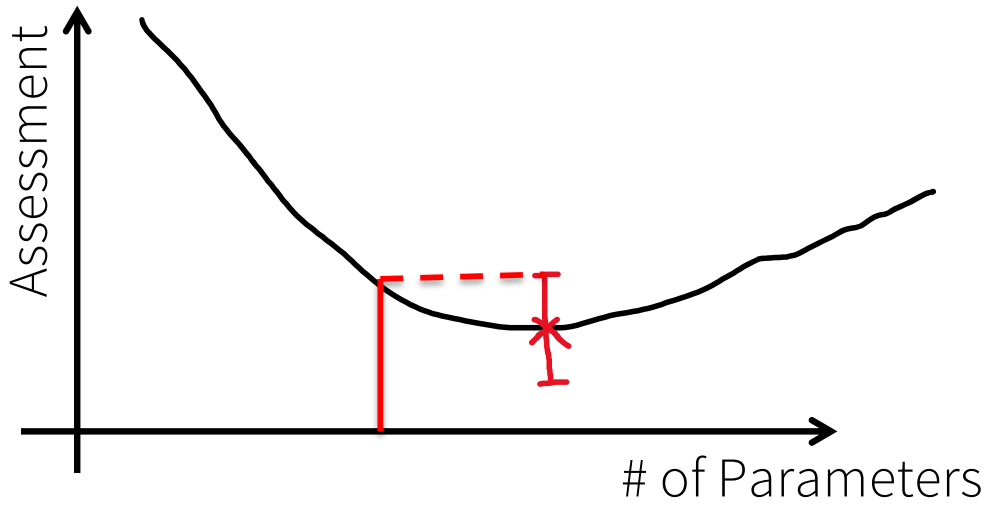
- The model assessment is done in three phases
  - Training
  - Test
  - Validation
- Why Validation

Sampling - Validation

- Cross validation or out-of-sample testing
- Leave-one out cross validation
- K-fold cross validation
- Bootstrap validation

Validation – One Standard-Error Rule

- The validation errors are calculated by randomly selecting training and validation set
- The end result will be average of many results with standard deviation of σ.
- The standard-Error rule says select simplest model within one standard deviation of the mean – which is the **smallest** model



Model Selection is applicable to many parameters, beyond linear regression

Model Selection

- Subset selection – identifying a subset of the parameters that is expected to give the best response
- Regularization / Shrinkage – automatic selection of the best coefficients of parameters by having a trade off between them

- Dimension reduction (Not covered in this course) – finding a best subset of dimensions achieving the best results, in which dimensions are combinations of parameters. Then projections in the new dimensions are used as predictors of the result.

Subset Selection

- Best subset selection with stepwise model. The goal is selecting the best *k* parameters out of total *p* parameters.

1. Predict output with the null model
2. For k=1,2, … , p
   1. Find all $\binom{p}{k}$ models
   2. Choose the best among $\binom{p}{k}$ models. The best can be used using R2 or Residual SS
3. Select the best model using adjusted R^2, or cross validation.

Stepwise Selection

- With large set of parameters, best subset selection is not feasible – Too large!

Forward Selection

1) Begin model with the null model.

   For example, start with simple mean

2) Try simple linear model by adding only one parameter and try all the models

3) Add a parameter with the lowest residual SS

4) Continue until a desirable accuracy reached based on a given rule

Forward Selection Details

1. Predict output with the null model
2. For k=1,2, … , p
   1. Find all $p - k$ models
   2. Choose the best among $p - k$ models. The best can be used using R^2 or Residual SS
3. Select the best model using adjusted R^2, or cross validation. (don't use R^2, since the model sizes are not the same)

Forward Stepwise

- Computationally very attractive
- Not optimum, but close to optimum

Backward Selection

1) Begin model with the all the parameters
2) Try a model by dropping only one parameter and try all the models
3) Drop a parameter which results to the highest residual SS
4) Continue until a desirable accuracy reached based on a given rule

Backward Selection Details

1. Predict output with the null model
2. For k=p, p-1, … 1
   1. Find all k models
   2. Choose the best among k models. The best can be used using R^2 or Residual SS
3. Select the best model using adjusted R^2, or cross validation. (don't use R^2, since the model sizes are not the same)

Note this is 1+p(p+1)/2 complexity compare to the total space : 2^p

Backward Stepwise

- Computationally very attractive
- Not optimum, but close to optimum
- Multi-polynomial doesn't have a solution if ($n<p$) number of observations is smaller than parameters.
  - Backward stepwise doesn't work for the same reason
  - Forward stepwise still work even ($n < p$)

Regularization

- L1 norm – Lasso
- L2 Norm – Ridge Regression
- Elastic net
- Li Norm

Regularization

- L2 Norm – Ridge Regression

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

Where $\lambda > 0$ is a tunning parameter.

- The first term is the standard sum square of error, which we tried to minimize.
- The second term is balancing against increasing weights, since more weights increases error.
- The tunning parameter $\lambda > 0$ controls forces of reducing number of parameters.
- The λ value has to be set by us! Cross validation is an effective way for this.

Ridge Regression – Scaling

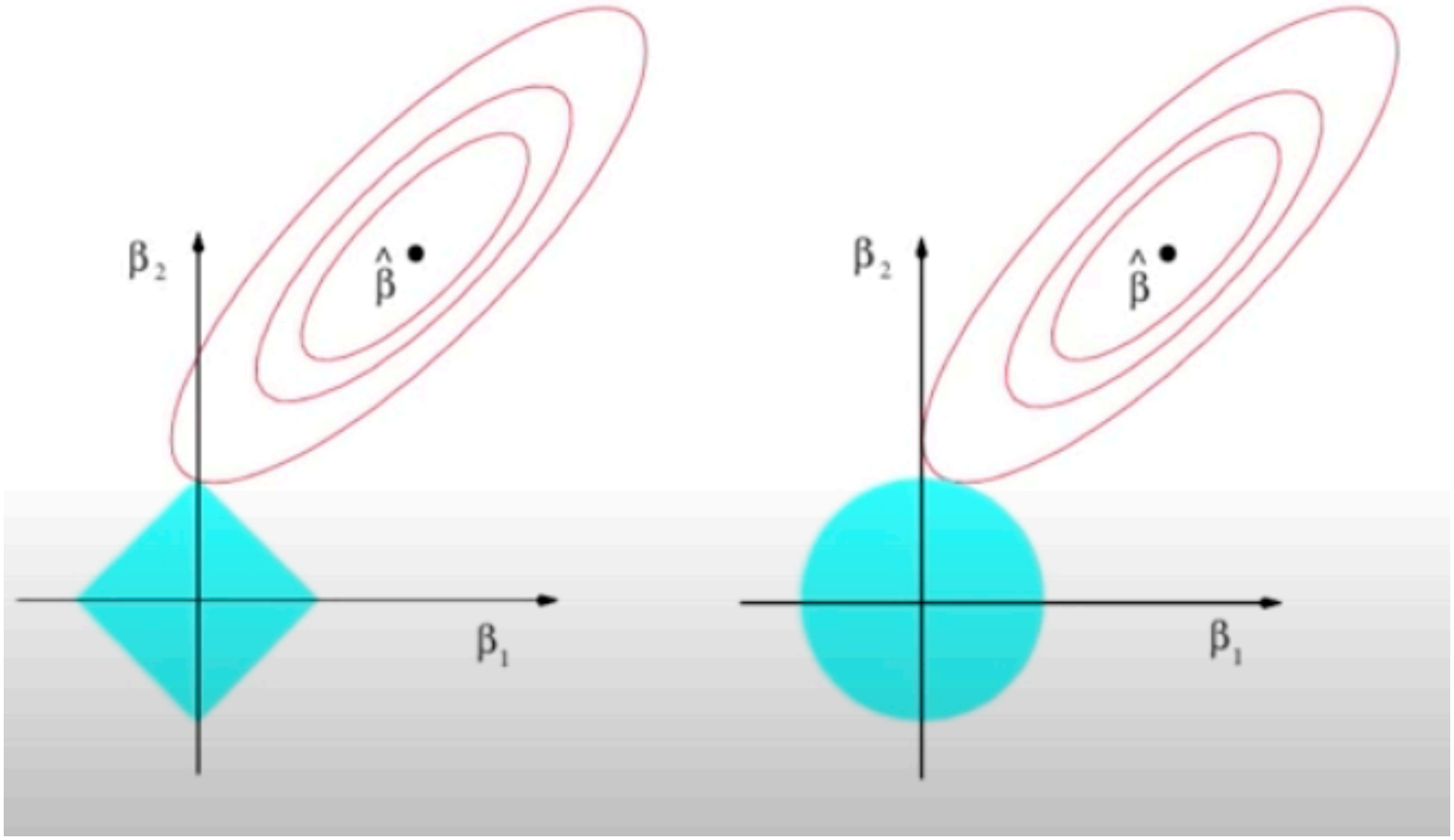
- Least square is equivariant: The standard least squares coefficients estimates are not dependent on scales
- BUT – ridge regression is sensitive to scale of the parameters.
- Therefore, parameters need to get standardized, e.g
- $$Standard\ Xi = \frac{X_i}{\sigma} = {X_i} \Big/ \sqrt{\frac{1}{n}\sum(X_i - \bar{X}_i)^2}$$

Lasso

- Ridge regression doesn't force small coefficients to zero
- Lasso is another option which doesn't have that issue

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

Lasso vs Ridge Regression



Elastic Net

- But Lasso forces too many small coefficient t become zero, which reduces the overall performance
- Elastic net combines Lasso and Ridge regression
  - Note that there is one more parameter which needs to be set!

Credit Balance Example

- Prediction credit balance as a function of 10 parameters including
  - Id
  - Income
  - Credit limit
  - Rating
  - Bing student of not
  - Number of credit cards
  - Age
  - Education
  - Gender
  - Married
  - Ethnicity
  - Balance

Reference: This example has been adopted from “an Intro to Statistical Learning” by G. James, et. Al.

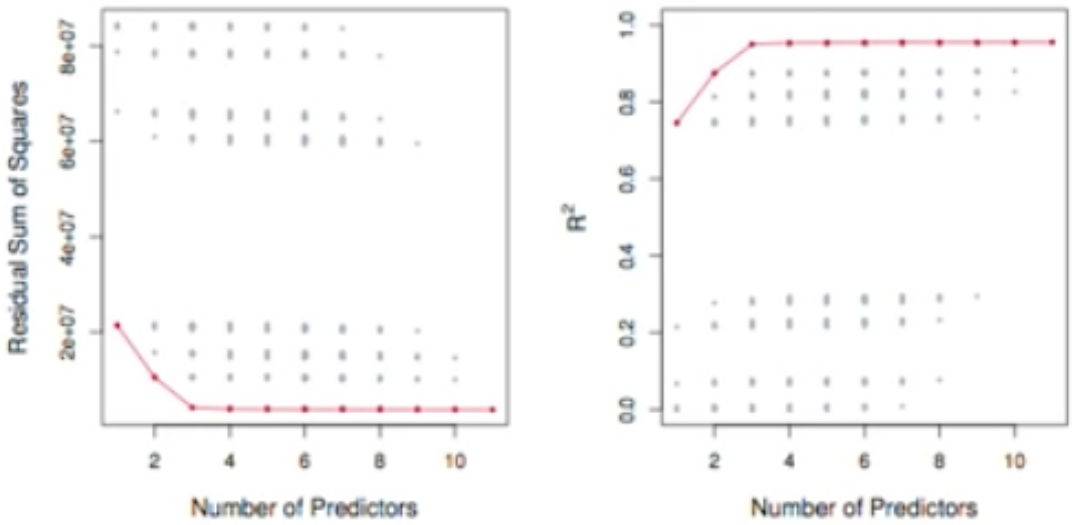
Credit Balance Example – Forward Selection

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income	rating, income,
	student, limit	student, limit

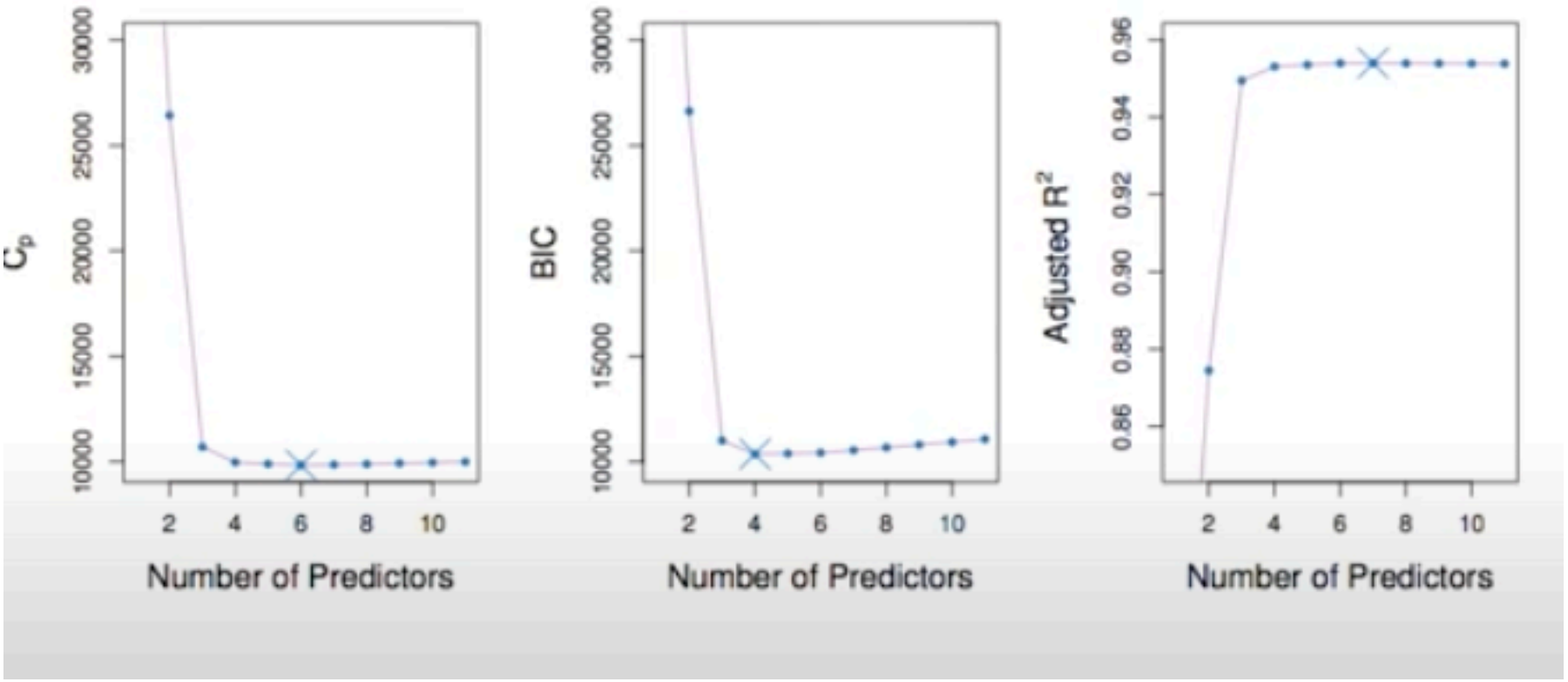
The first four selected models for best subset selection and forward stepwise selection on the Credit data set. The first three models are identical but the fourth models differ.

Credit Balance Example - Number of Predictors

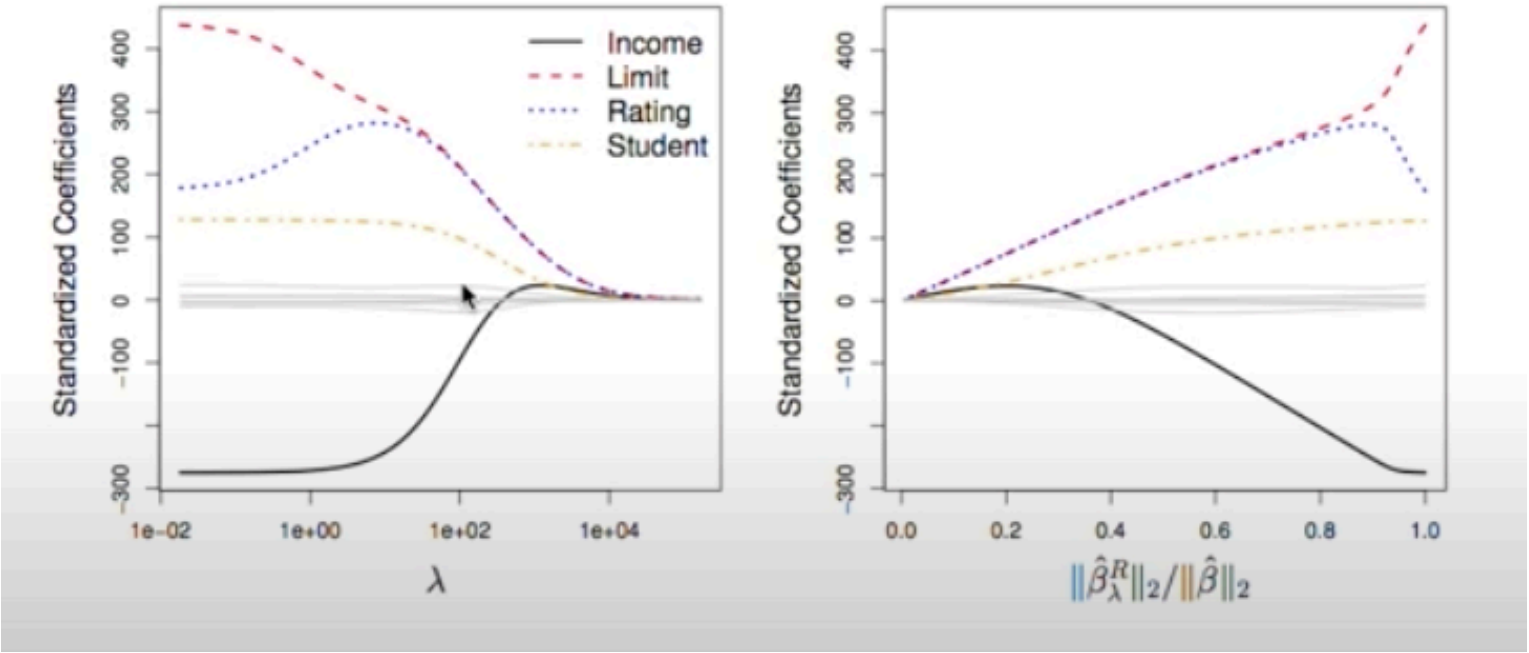
- Number of predictors impact on error



Credit Balance Example – Assessment of Parameters



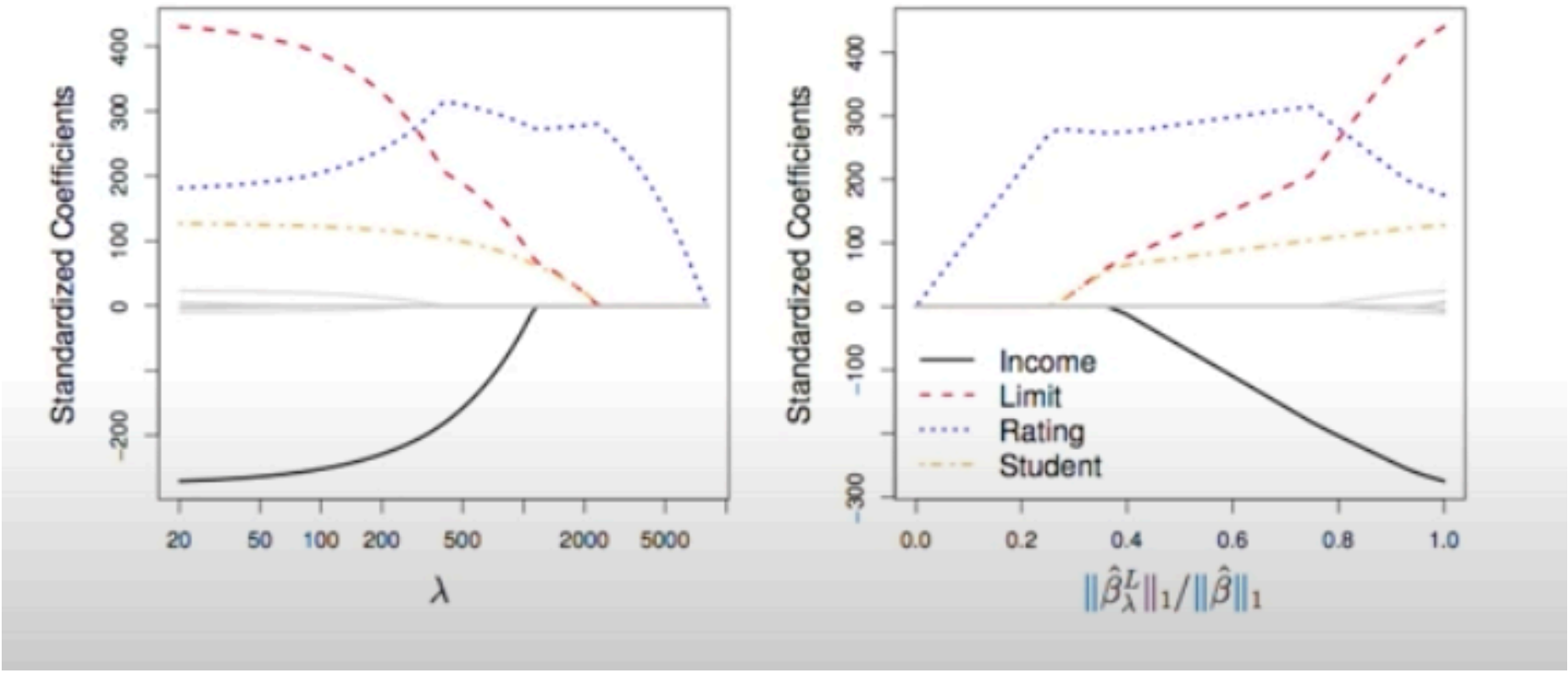
Example Credit Card - Ridge Regression



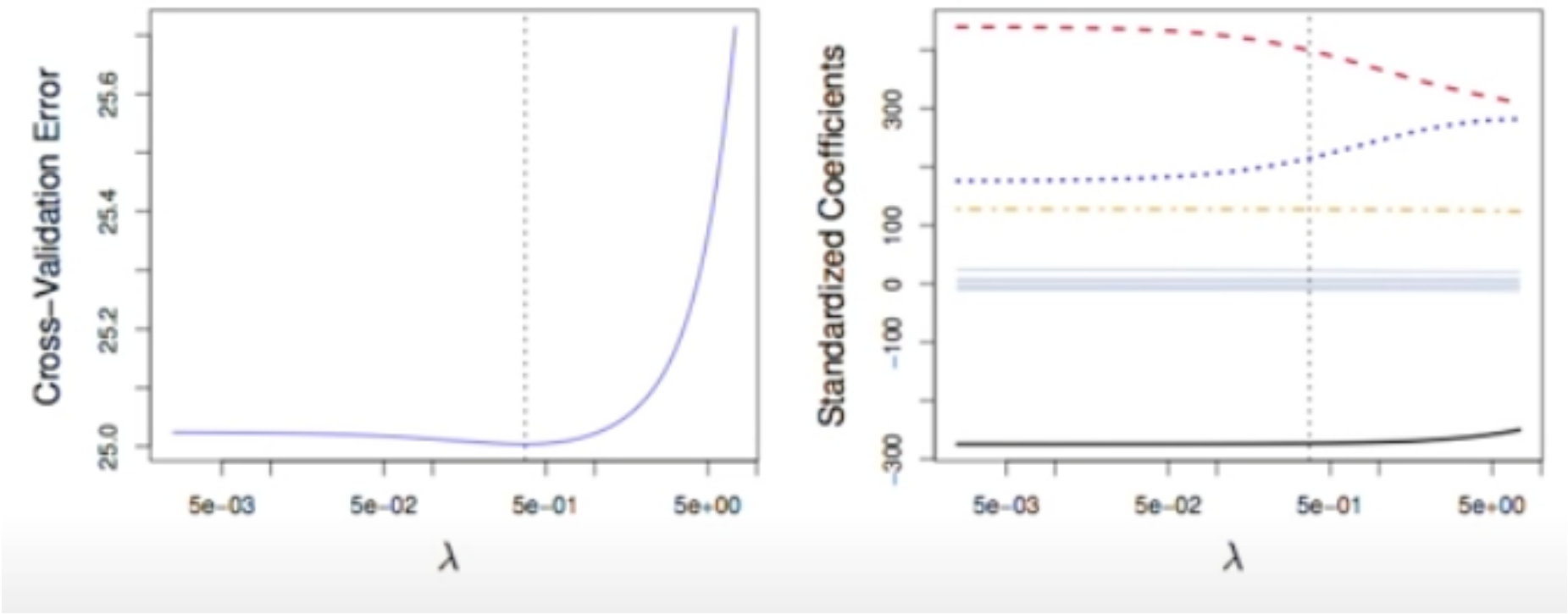
- Note some coefficients are close to zero but not zero. That can be an issue with Ridge regression.

Example - Lasso

Notice weight changes



Credit Card Example – Ridge Parameter Selection



Credit Card Example – Cross Validation

- Validation – ¾ data vs ¼ training
- 10 fold Cross validation
- BIC as expected learns toward smaller model