Lingnan University 嶺南大學 香港 Hong Kong

School of Data Science 數據科學學院

CDS 533
Statistics for Data Science

Instructor: Lisha Yu

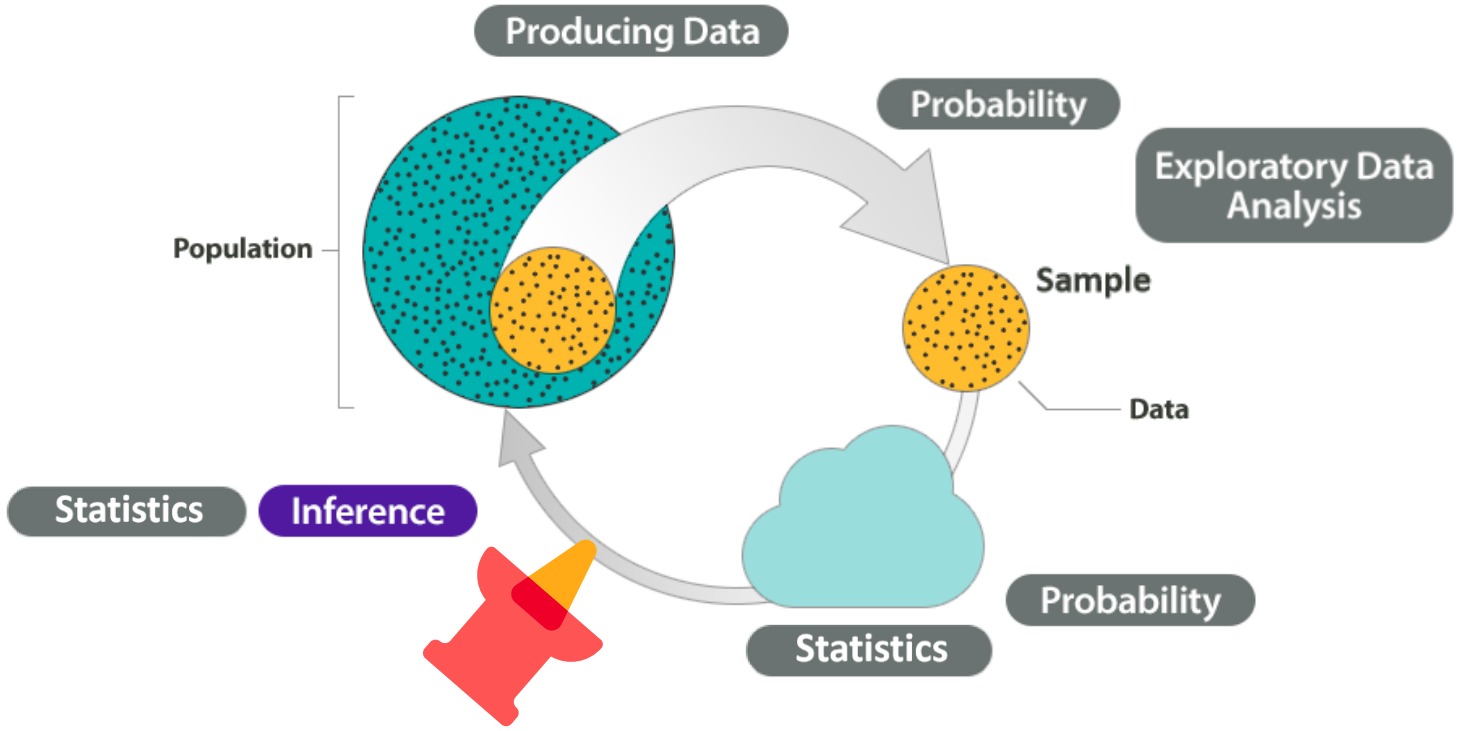
Division of Artificial Intelligence

School of Data Science

Lingnan University

Fall 2024

Big Picture of Statistics



Statistical Inference
(Model Selection and Model Regularization)

Recap: K-fold Cross-validation

Common choices of K: $K = 5$ or $K = 10$

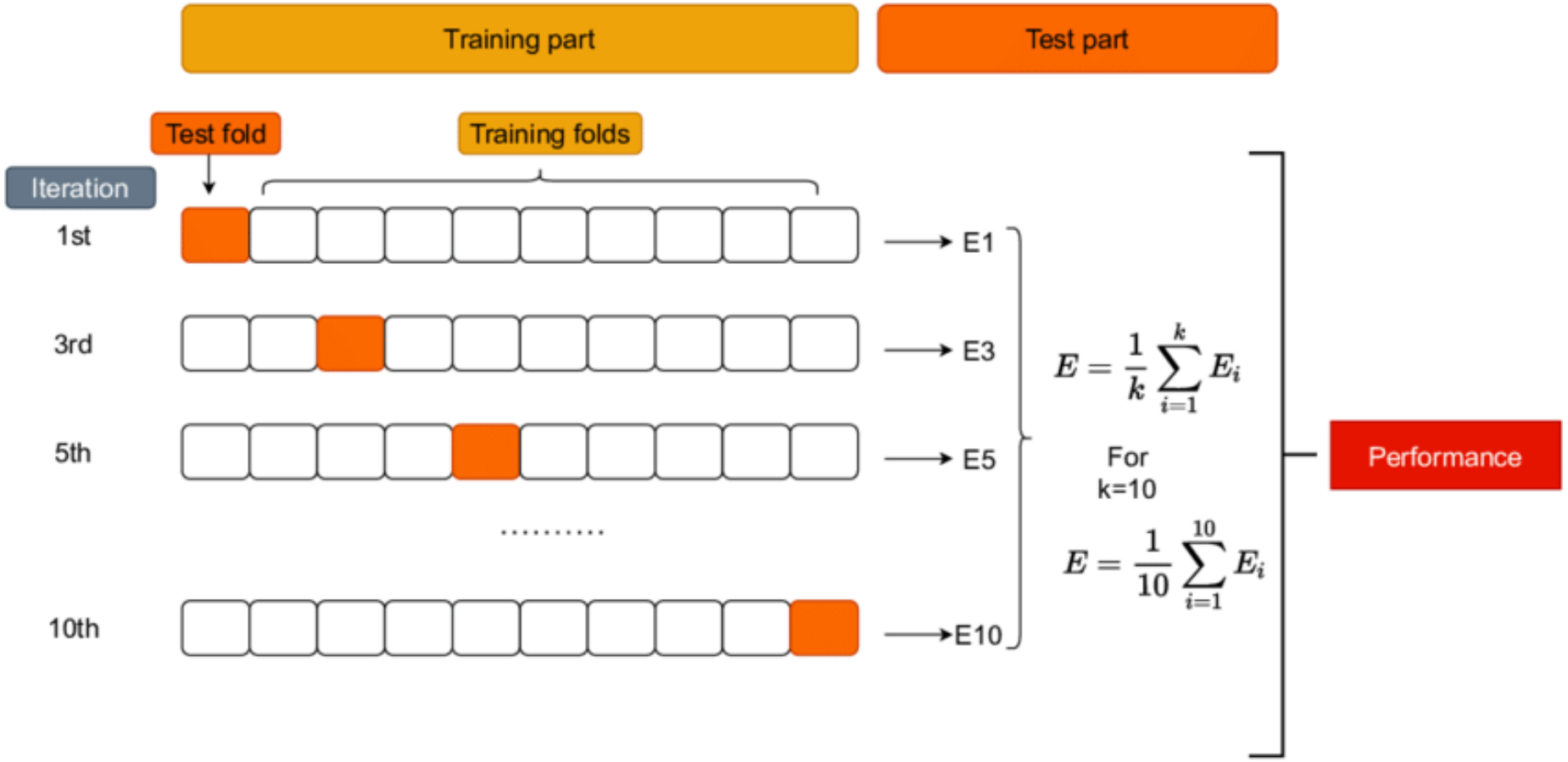
- Advantage over LOOCV:
  1. Computationally lighter, especially for complex models with large data.
  2. Offers a good balance between bias and variance in model performance estimates.

  Advantage over validation set approach:
  1. Less variability resulting from the data-split, thanks to the averaging.

K-FOLD CROSS-VALIDATION	Easy to implement, computationally efficient	Can be biased towards majority classes in imbalanced datasets
STRATIFIED CROSS-VALIDATION	Reduces bias towards majority classes in imbalanced datasets	Can be computationally expensive for large datasets
LOOCV	Provides a very accurate estimate of the model's generalization performance	Very computationally expensive, can induce bias

Recap: K-fold Cross-validation

Repeated K-fold Cross-validation



Performance Metrics for Regression

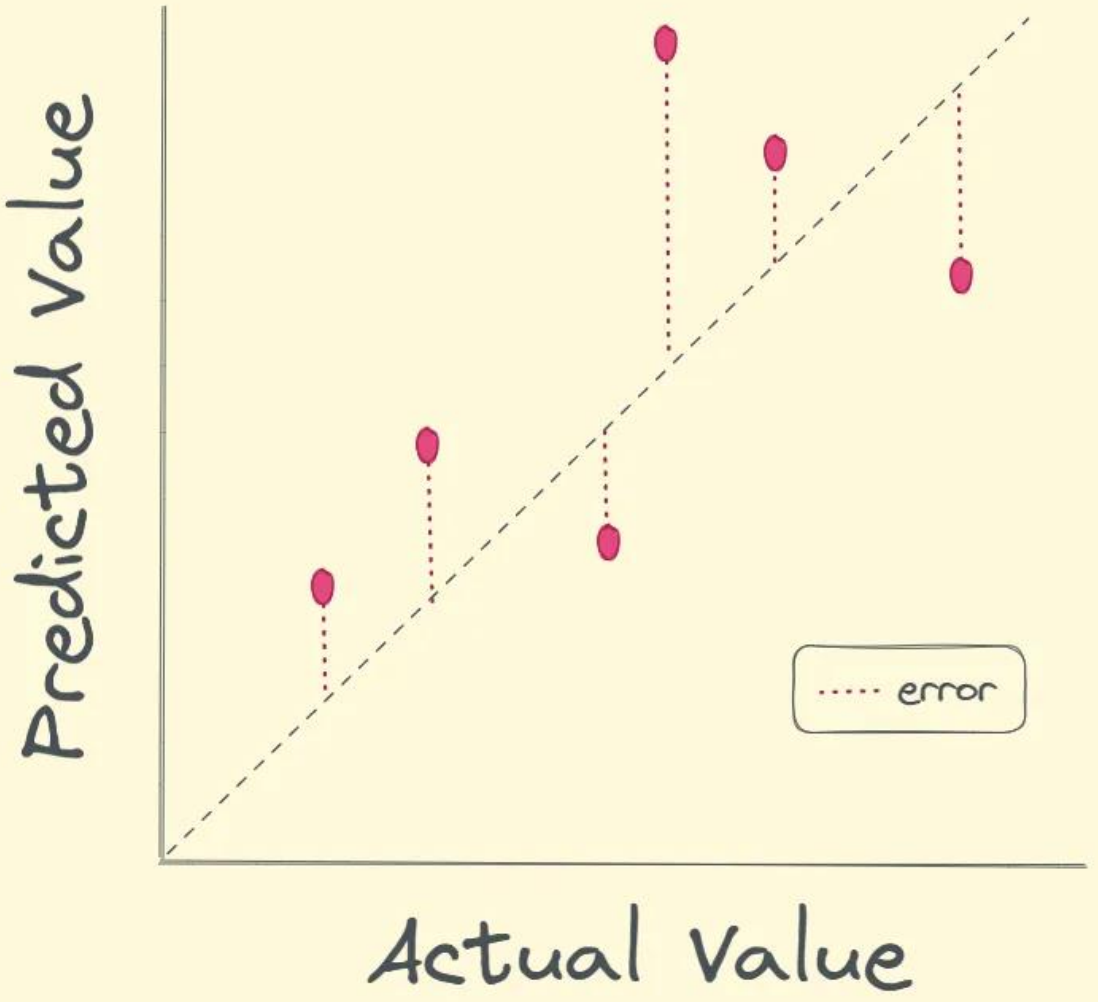
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$



Performance Metrics for Regression

Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- higher values → more significant discrepancies between predicted and actual values.
- MSE is sensitive to outliers; commonly used due to its mathematical ;less interpretable.

Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

- RMSE is in the same units as the dependent variable, making it more interpretable.
- RMSE is preferred when the distribution of errors is not normal or when outliers are present, as it mitigates the impact of large errors.

Performance Metrics for Regression

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

- MAE is less sensitive to outliers than MSE but may not adequately penalize large errors.

Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{|Y_i|}$$

- MAPE is suitable for assessing relative errors and comparing in percentage terms

R-squared (R^2)

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y_i})^2}$$

Practical Interpretation and Applications

Let’s explore some scenario-based examples to understand how these metrics are utilized:

Example 1: Predicting Housing Prices

Scenario: A real estate agency wants to build a predictive model to estimate housing prices based on location, square footage, number of bedrooms, and amenities.

Regression Metric: Mean Squared Error (MSE) or Root Mean Squared Error (RMSE).

Practical Interpretation and Applications

Let's explore some scenario-based examples to understand how these metrics are utilized:

Example 2: Forecasting Sales

Scenario: A retail chain wants to forecast store sales based on historical data, including promotional activities, seasonality, and economic indicators.

Regression Metric: Mean Absolute Error (MAE)

Practical Interpretation and Applications

Let's explore some scenario-based examples to understand how these metrics are utilized:

Example 3: Predicting Crop Yields

Scenario: An agricultural research institute aims to predict crop yields based on soil quality, weather conditions, irrigation, and crop varieties.

Regression Metric: R-squared (R^2).

Practical Interpretation and Applications

Let's explore some scenario-based examples to understand how these metrics are utilized:

Example 4: Estimating Customer Lifetime Value (CLV)

Scenario: A subscription-based business wants to estimate the Customer Lifetime Value (CLV) to optimize marketing strategies and improve customer retention.

Regression Metric: Mean Absolute Percentage Error (MAPE).

How to Choose Appropriate Performance

1. Forecasting with Outliers

Scenario: Predicting sales figures for a retail store with occasional outlier events, such as seasonal promotions or product launches.

Regression Metric: Mean Absolute Percentage Error (MAPE).

Rationale: MAE is less sensitive to outliers than MSE or RMSE, making it suitable for scenarios where occasional extreme values may distort the accuracy assessment.

How to Choose Appropriate Performance

2. Model Interpretability

Scenario: Developing a regression model to predict housing prices, where stakeholders prioritize interpretability and ease of understanding.

Regression Metric: R-squared (R^2).

Rationale: R^2 provides a straightforward interpretation of the proportion of variance explained by the model, facilitating communication and decision-making among stakeholders.

How to Choose Appropriate Performance

3. Relative Error Consideration

Scenario: Estimating customer churn rates for a subscription-based service, where relative errors are more critical than absolute errors.

Regression Metric: Mean Absolute Percentage Error (MAPE).

Rationale: MAPE expresses prediction accuracy in percentage terms, making it suitable for assessing relative errors and comparing performance across different datasets.

Linear Regression

Recall in the regression setting, the standard linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

is commonly used to describe the relationship between a response Y and a set of variables $X_1, X_2, \ldots, X_p$.

In Praise of Linear Models

The linear model has distinct advantages in terms of **inference** and is often shows good **predictive performance**.

- Some ways in which the simple linear model can be improved, by replacing plain least squares fitting with some alternative fitting procedures.

Why consider alternatives to least squares?

- Prediction Accuracy
  - especially when $p > n$, to control the variance.
- Model Interpretability
  - by removing irrelevant features, that is, by setting the corresponding coefficient estimates to zero, we can obtain a model that is more easily interpreted.

Three Classes of Methods

- **Subset Selection**. We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.

- **Shrinkage.** We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as regularization) reduces variance and can also perform variable selection.

- **Dimension Reduction**. We project the p predictors into a M-dimensional subspace, where M < p. [e.g., Principle Components Analysis]

Subset Selection

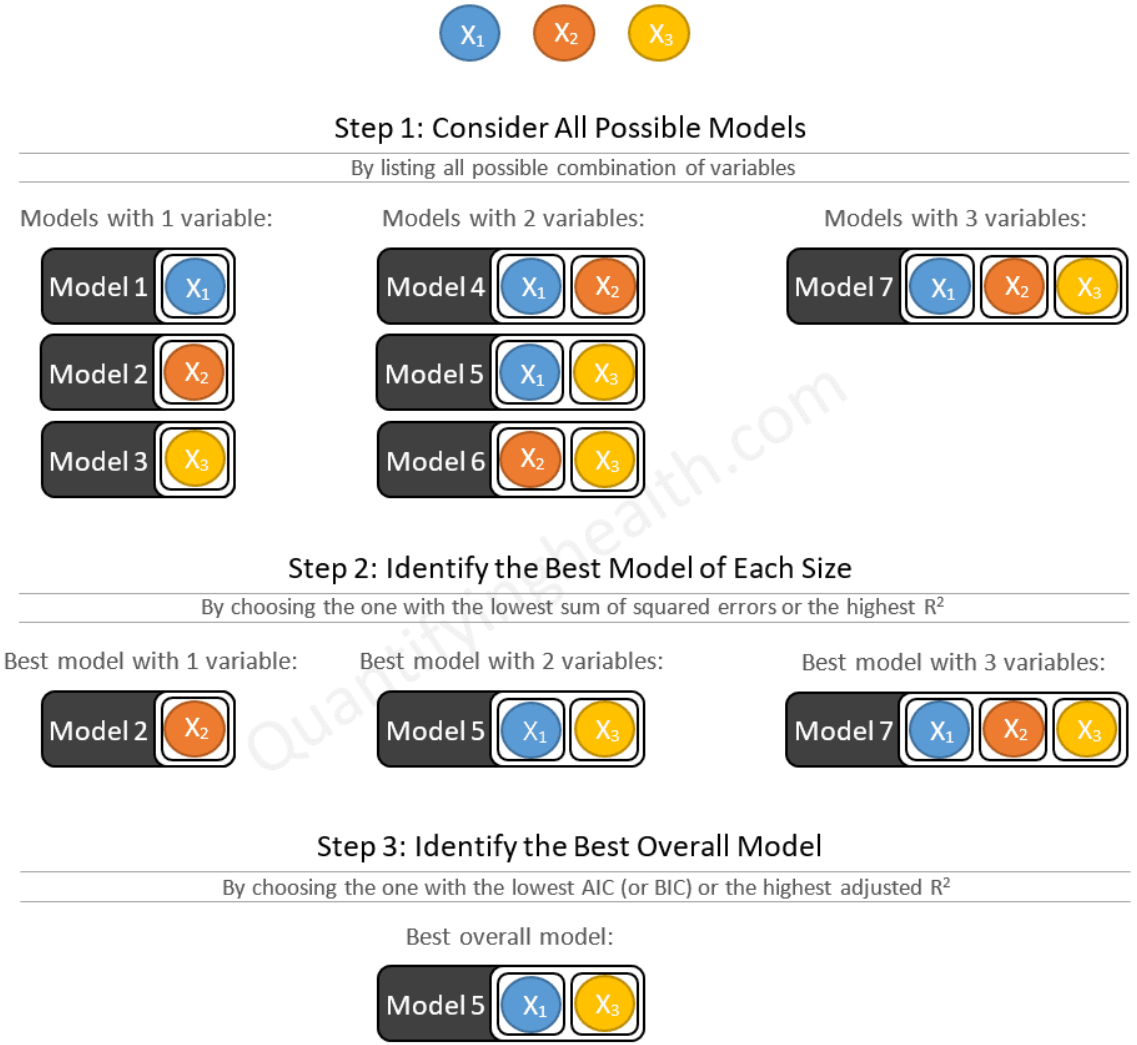
Best subset selection is a method that aims to find the subset of independent variables (X_i) that best predict the response (Y) and it does so by considering all possible combinations of independent variables.

How best subset selection works?

1. Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   b) Pick the best among these $\binom{p}{k}$ models, and call it M_k . Here **best** is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $M_0, \ldots, Mp$ using cross-validated prediction error, Cp, (AIC), BIC, or adjusted R^2.

Subset Selection

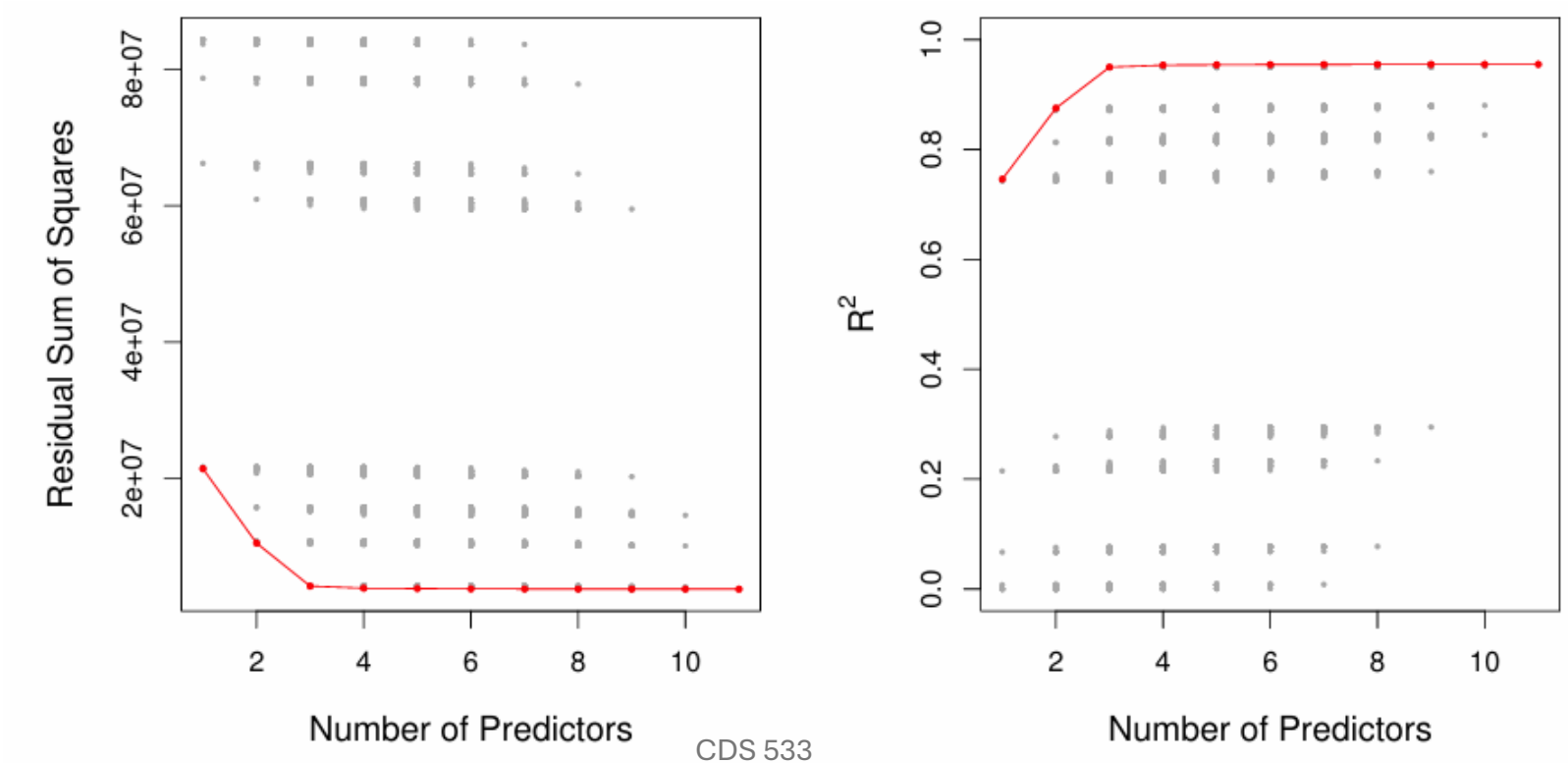
- **Best subset selection:** it ends up selecting 1 model from 2^p possible models.



Credit Dataset

Response: balance (average credit card debt)

Predictors: age, **cards** (number of credit cards), **education** (years of education), **income** (in dollars), **limit** (credit limit), **rating** (credit rating), **gender**, **student** (Yes/No), **marriage status**, **ethnicity**



Advantages and Limitations

Advantages:

- Yields a simple and easily interpretable model
- Provides a reproducible and objective way to reduce the number of predictors compared to manually choosing variables

Limitations:

- Computational limitation
  - e.g., for 3 predictors, consider 2^3 = 8 models; for 10 predictors, 2^{10} = 1024 models
- Best subset selection may also suffer from an enormous search space can lead to overfitting and high variance of the coefficient estimates.

Stepwise Selection (Forward)

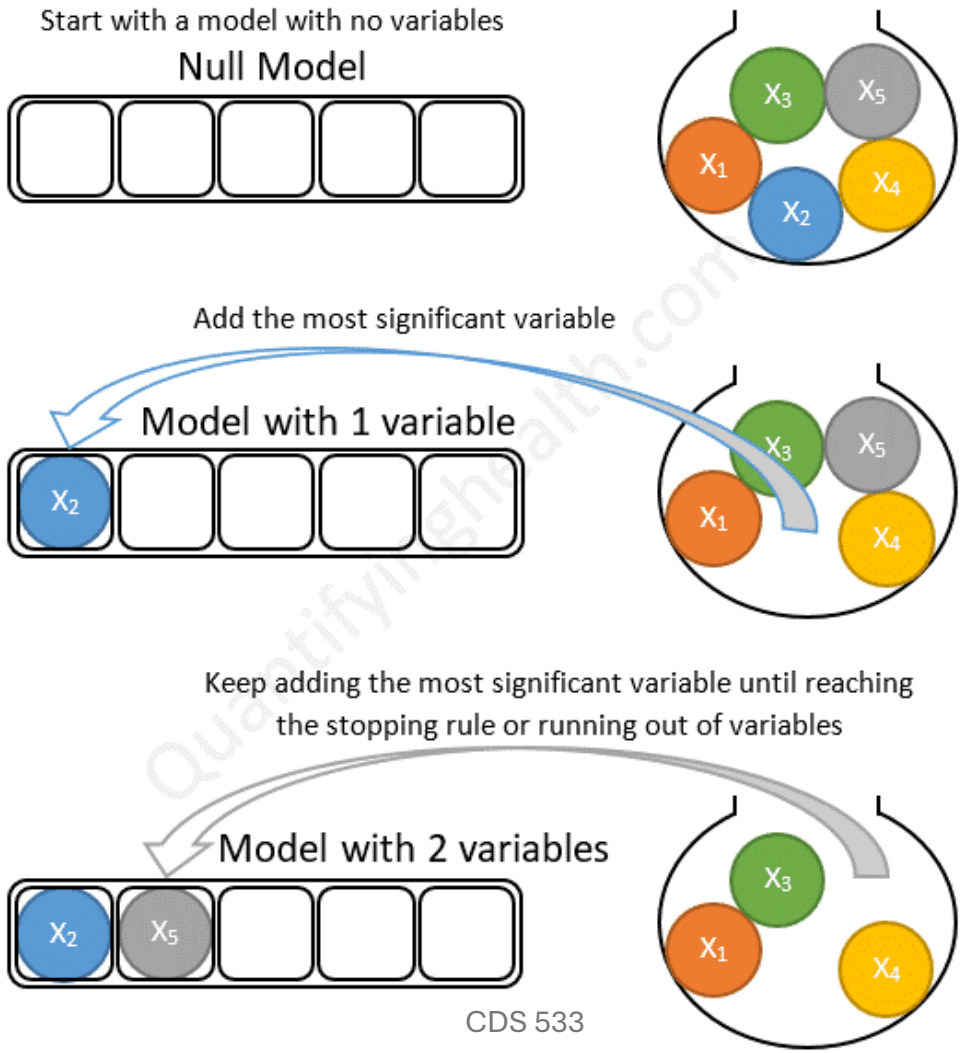
Forward stepwise selection (or forward selection) begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. In particular, at each step, the variable that gives the greatest additional improvement to the fit is added to the model.

How forward selection works?

1. Let M_0 denote the null model, which contains no predictors.
2. For $k = 0, 1, 2, \ldots p-1$:
   a) Consider all $p-k$ models that augment the predictors in M_k with one additional predictor.
   b) Choose the best among these $p-k$ models, and call it M_{k+1} .Here **best** is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $M_0, \ldots, Mp$ using cross-validated prediction error, Cp, (AIC), BIC, or adjusted R^2.

Forward Selection

- **Forward selection:** example with 5 variables



More on Forward Selection

- Computational advantage over best subset selection is clear.
- It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

# 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

Response: balance (average credit card debt)

Predictors: age, **cards** (number of credit cards), **education** (years of education), **income** (in dollars), **limit** (credit limit), **rating** (credit rating), **gender**, **student** (Yes/No), **marriage status**, **ethnicity**

Stepwise Selection (Backward)

- Like forward stepwise selection, **backward stepwise selection** provides an efficient alternative to best subset selection.
- However, unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

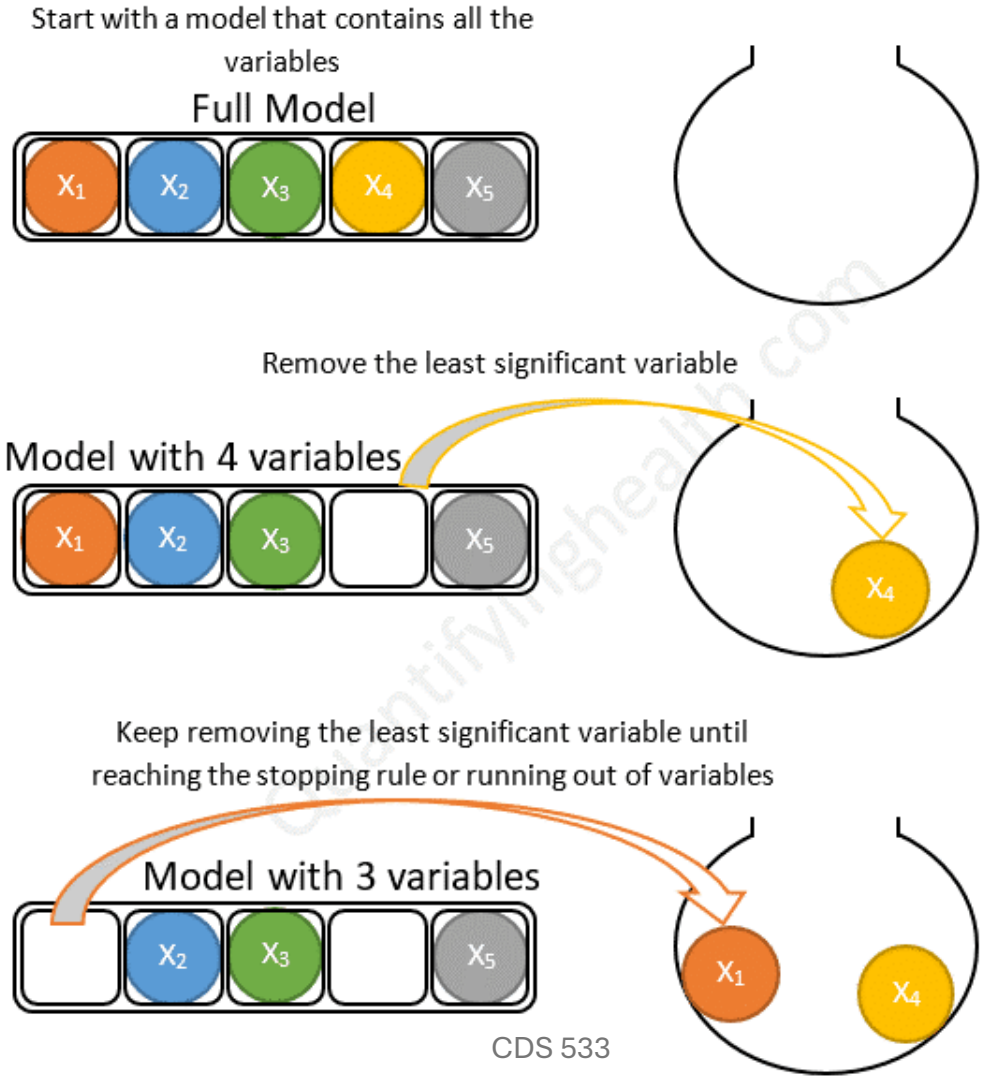
Stepwise Selection (Backward)

How backward selection works?

1. Let M_p denote the full model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 2, 1$:
   a) Consider all k models that contain all but one of the predictors in M_k, for a total of $k-1$ predictors.
   b) Choose the best among these k models, and call it M_{k-1} .Here **best** is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $M_0, \ldots, Mp$ using cross-validated prediction error, Cp, (AIC), BIC, or adjusted R^2.

Backward Selection

- **Backward selection:** example with 5 variables



More on Backward Selection

- Like forward stepwise selection, the backward selection approach searches through only $1 + p(p + 1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection
- Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the **best** model containing a subset of the p predictors.
- Backward selection requires that the number of samples n is larger than the number of variables p (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.

Choosing the Optimal Model

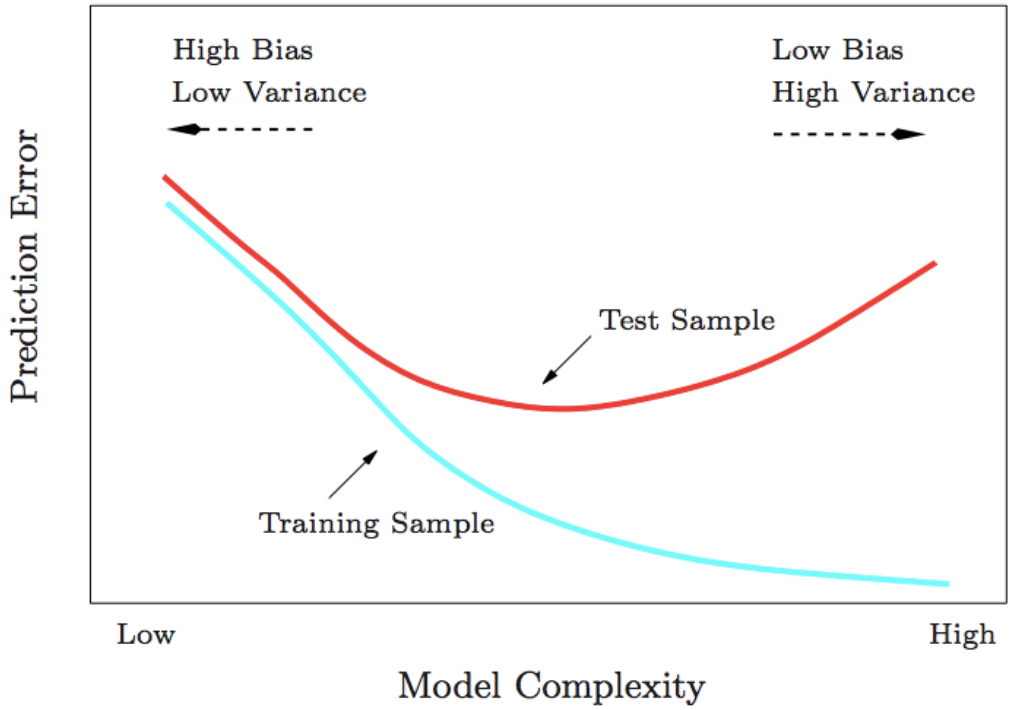
- Unlike step 2, where we compared models of the same size, in step 3 we will be comparing models of **different sizes**, so we can no longer use RSS or R^2 to select the best overall model.

Why?

- Because the model with most variables always have the lowest RSS and the highest R^2 .
- One solution is to choose the best overall model according to a statistic that imposes some sort of penalty on bigger models, especially when they contain additional variables that barely provide any improvement.

Choosing the Optimal Model

- We wish to choose a model with low test error, not a model with low training error. Recall that training error is usually a poor estimate of test error.
- Therefore, RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.



Estimating Test Error: Two Approaches

1. We can indirectly estimate test error by making an **adjustment** to the training error to account for the bias due to overfitting.
2. We can **directly** estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in previous lectures.

- We illustrate both approaches next.

C_p, AIC, BIC, and Adjusted R^2

- These techniques adjust the training error for the model size, and can be used to select among a set of models with different numbers of variables.
- The next figure displays C_p, BIC, and adjusted R^2 for the best model of each size produced by best subset selection on the Credit data set.

Details on C_p and AIC

- **Mallow's C_p**

$$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right),$$

where d is the total # of parameters used and σ^2 is an estimate of the variance of the error $\in$ associated with each response measurement.

- **Akaike information criterion (AIC)**
  - The AIC criterion is defined for a large class of models fit by maximum likelihood:

$$\mathrm{AIC} = -2\log L + 2 \cdot d$$

where L is the maximized value of the likelihood function for the estimated model.

- In the case of the linear model with Gaussian errors, maximum likelihood and least squares are the same thing, and $\boldsymbol{C_p}$ and AIC are equivalent.

Details on BIC and Adjusted R^2

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n)d\hat{\sigma}^2\right).$$

- Like Cp, the BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value.
- BIC replaces the $2d\sigma^2$ in C_p with a $\log(n)d\sigma^2$ term, where n is the # of observations; but heavier penalty on models with many variables ($\log n$>2 for any $n > 7$).

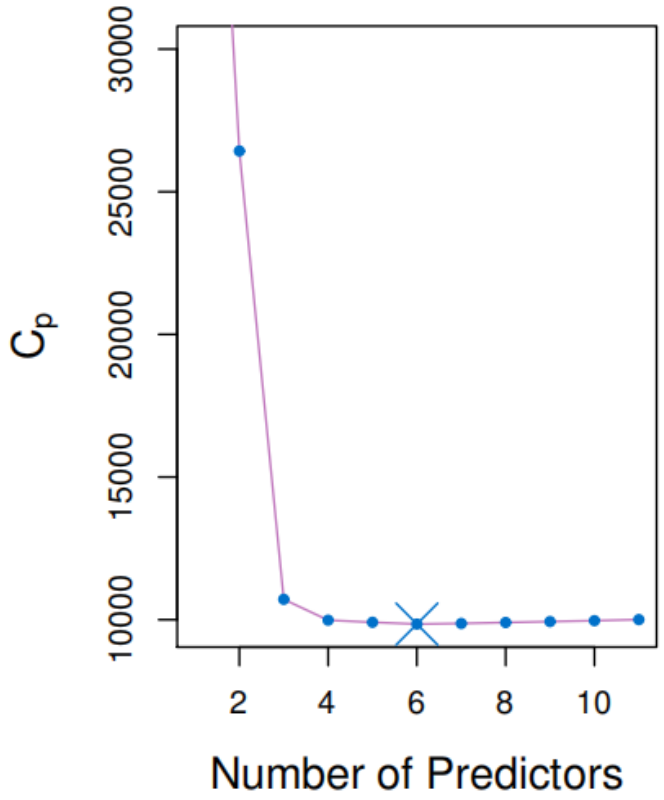
- **Adjusted R^2**
  - For a least squares model with d variables, the adjusted R^2 statistic is calculated as

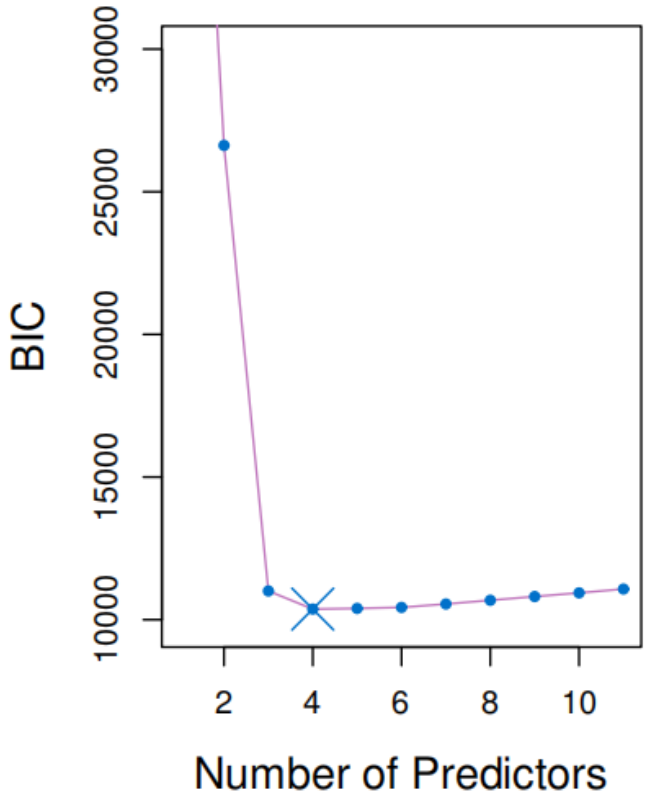
$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}.$$

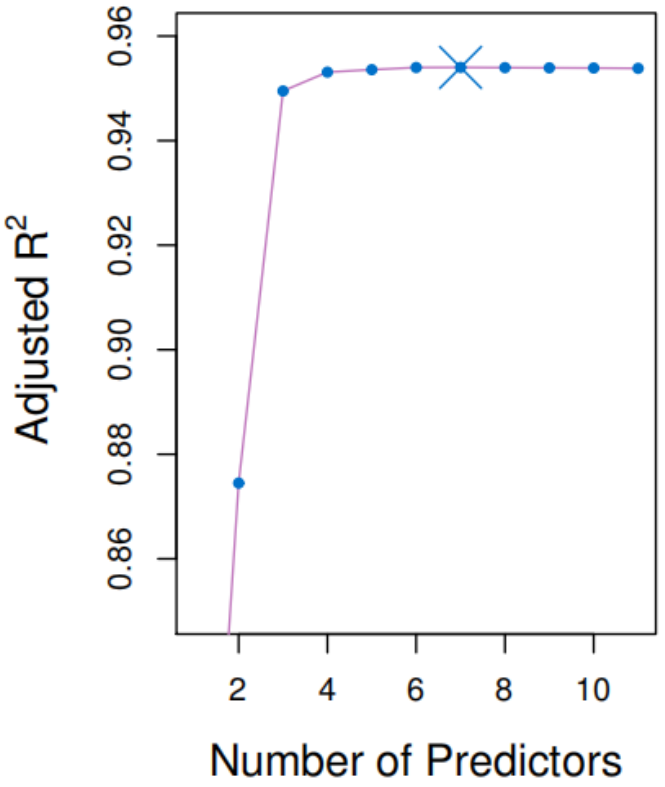
  - Unlike C_p, AIC, BIC, for which a small value indicates a model with a low test error, a large value of adjusted R^2 indicates a model with a small test error.

C_p, AIC, BIC, and Adjusted R^2

- **Credit data example**







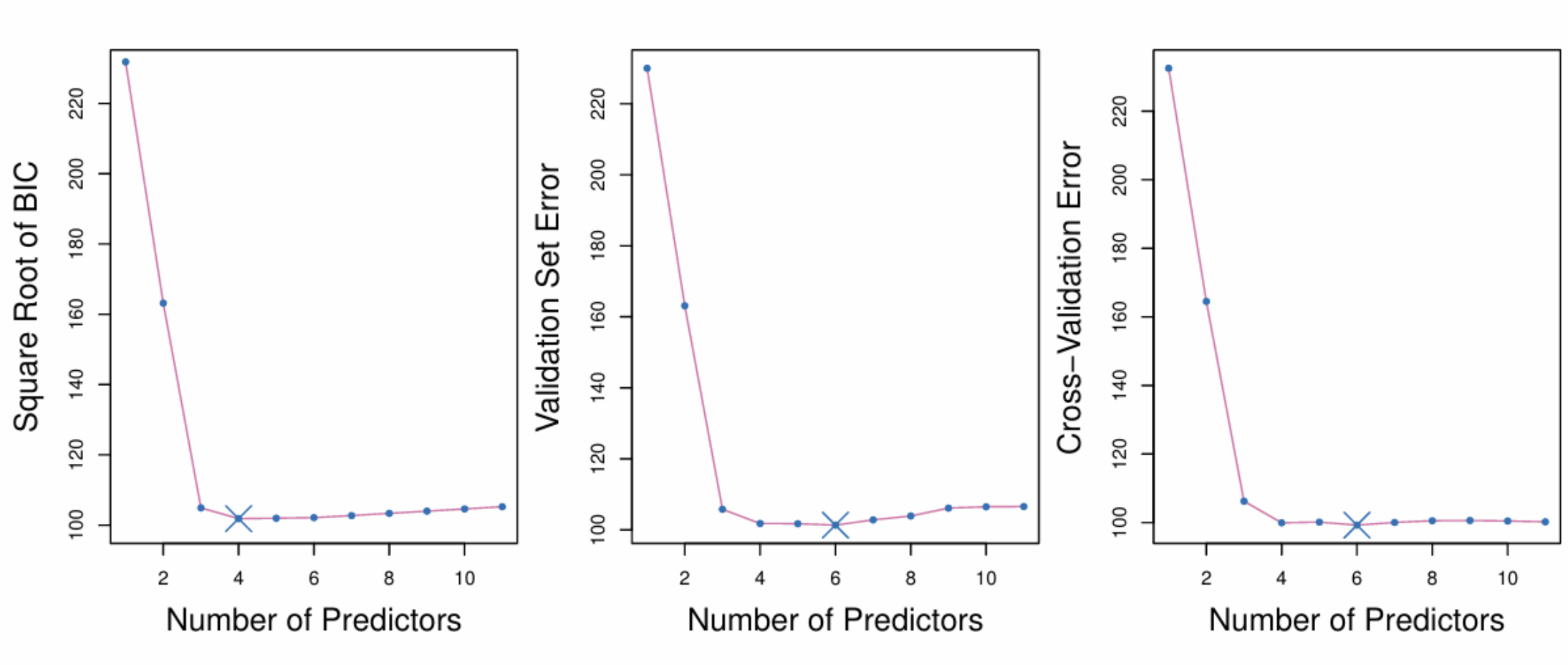
Validation and Cross-Validation

In Step 2, each of the procedures returns a sequence of models M_k indexed by model size $k = 0, 1, 2, \ldots$. Our job here is to select k. Once selected, we will return model M_k.

- We compute the validation set error or the cross-validation error for each model M_k under consideration and then select the k for which the resulting estimated test error is smallest.
- This procedure has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2.

Validation and Cross-Validation

Credit dataset

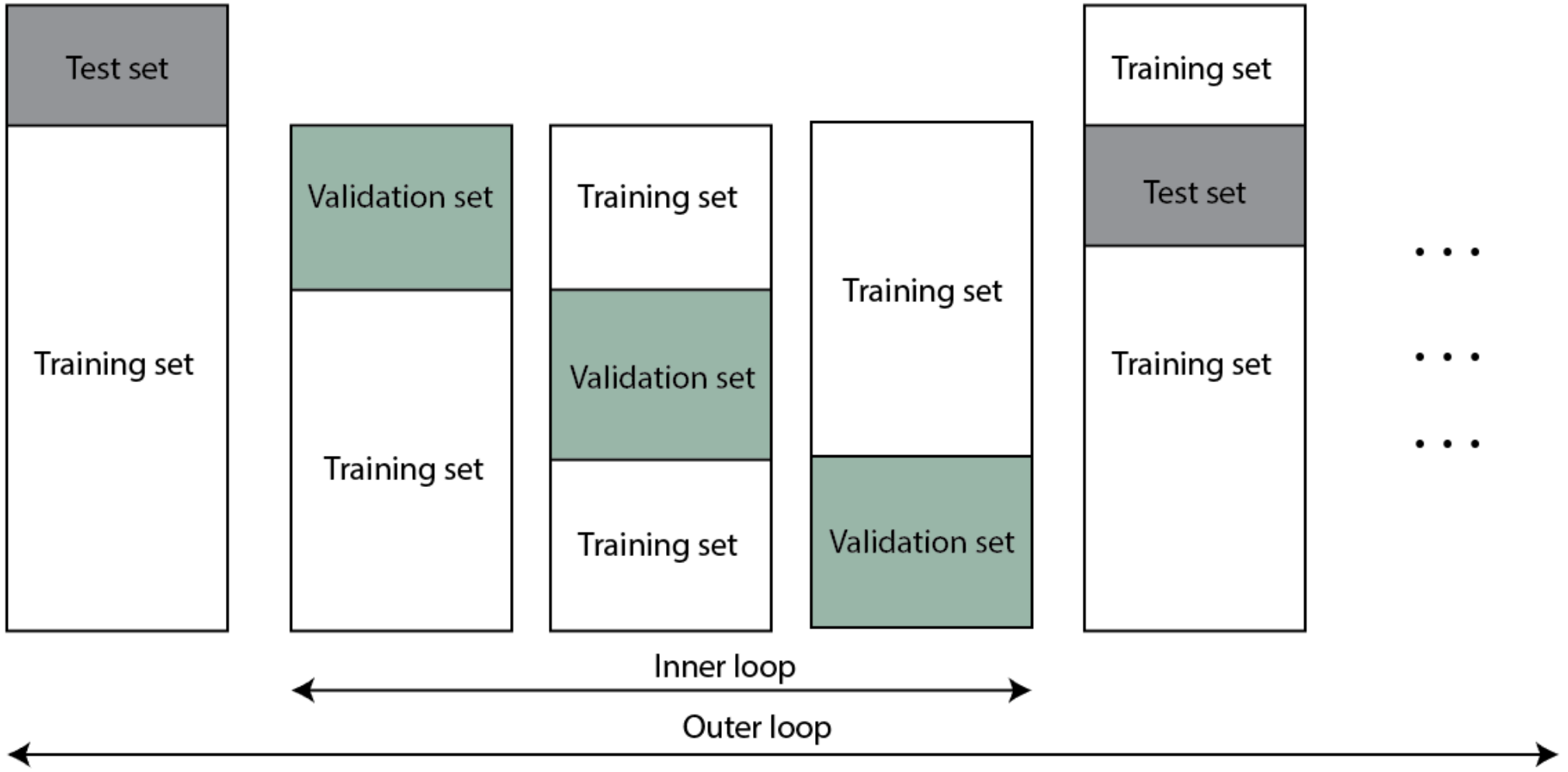


Details of Previous Figure

- The **validation errors** were calculated by randomly selecting 3/4 of the observations as the training set, and the remainder as the validation set.
- The **cross-validation errors** were computed using k = 10 folds. In this case, the validation and cross-validation methods both result in a six-variable model.
- However, all three approaches suggest that the four-, ve-, and six-variable models are roughly equivalent in terms of their test errors.

- In this setting, we can select a model using the **one-standard-error rule**. We first calculate the standard error of the estimated test MSE for each model size, and then select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve.

More on Cross-validation Error



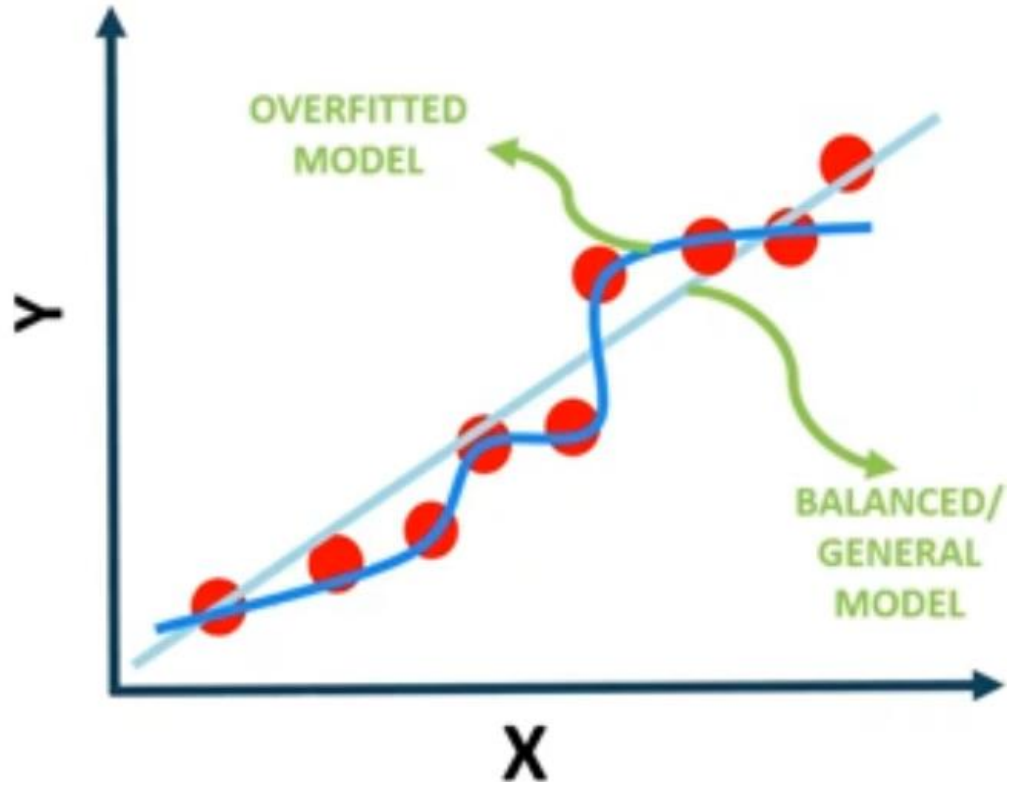
Shrinkage Methods

Ridge regression and Lasso

- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors.
- As an alternative, we can fit a model containing all p predictors using a technique that **constrains** or **regularizes** the coefficient estimates or, equivalently, shrinks the coefficient estimates towards zero.

Ridge Regression (L2 Regularization)

- Ridge regression advantage is to **avoid overfitting**.
- Our ultimate model is the one that could generalize patterns, i.e., works best on the training and testing dataset



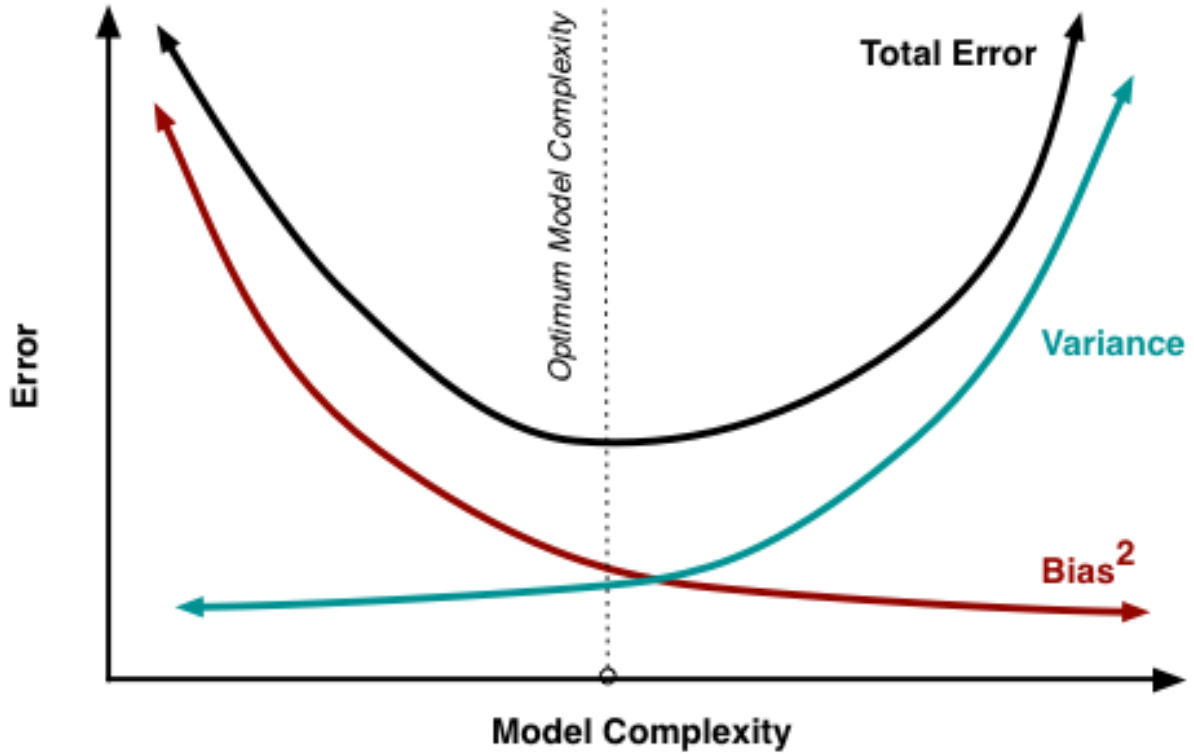
Ridge Regression (L2 Regularization)

- **Overfitting** occurs when the trained model performs well on the training data and performs poorly on the testing datasets

- Least sum of squares is applied to obtain the best fit line
- Since the line passes through the 3 training dataset points, the RSS = 0
- However, for the testing dataset, the RSS is large so the line has high variance
- Variance means that there is a difference in fit (or variability) between the training dataset and testing dataset

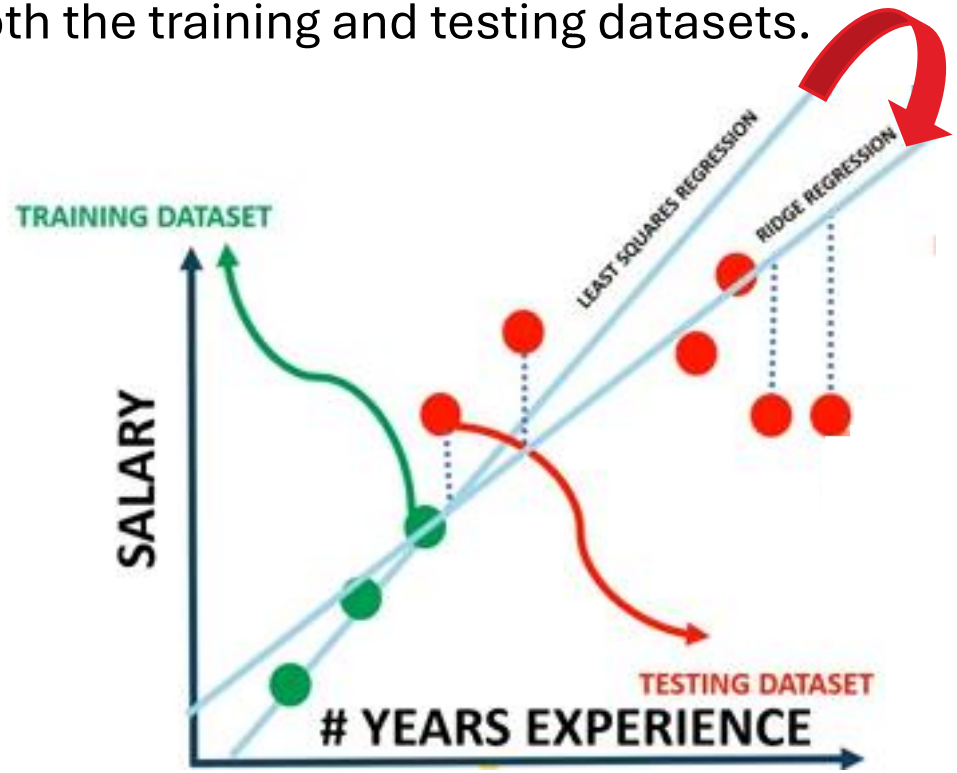
Ridge Regression (L2 Regularization)

- Ridge regression works by applying a **penalizing term** (reducing the weights and biases) to overcome overfitting.



Ridge Regression (L2 Regularization)

- Ridge regression works by attempting to increase the bias to improve variance (generalization capability)
- This works by changing the slope of the line
- The model performance might be a little poor on the training set, but it will perform consistently well on both the training and testing datasets.



Ridge Regression (L2 Regularization)

- Slope has been reduced with ridge regression penalty and therefore the model becomes **less sensitive** to changes in the independent variable (#Year of experience)

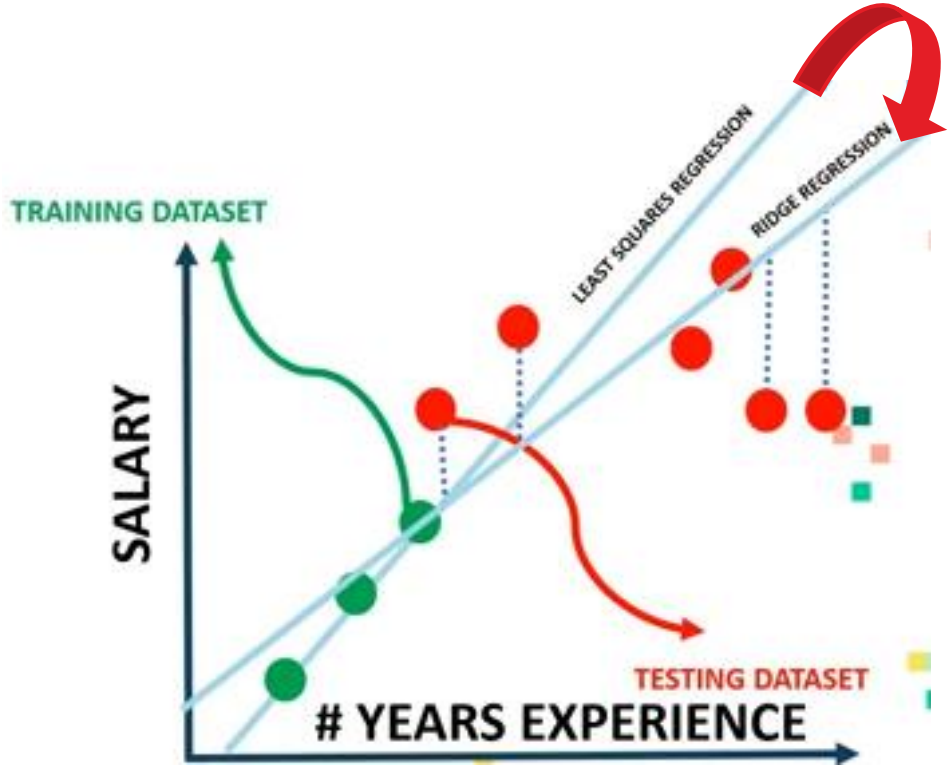
Least Square Regression:

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

Ridge Regression

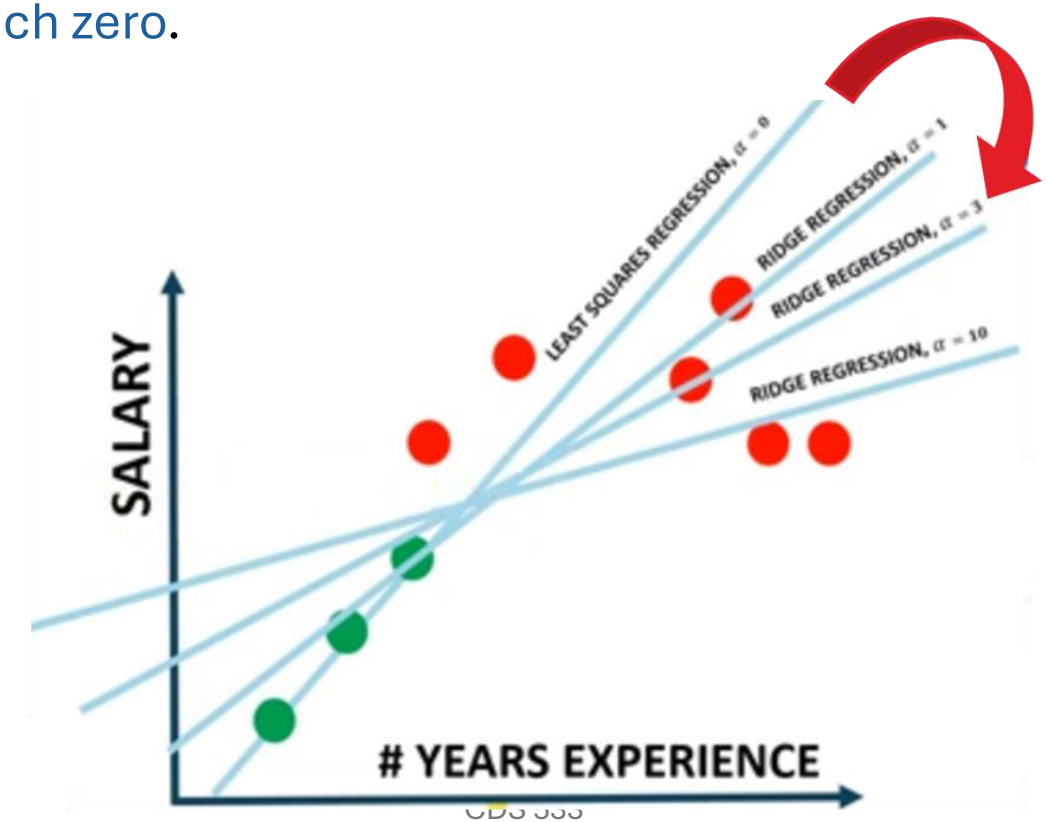
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

SHRINKAGE PENALTY TERM



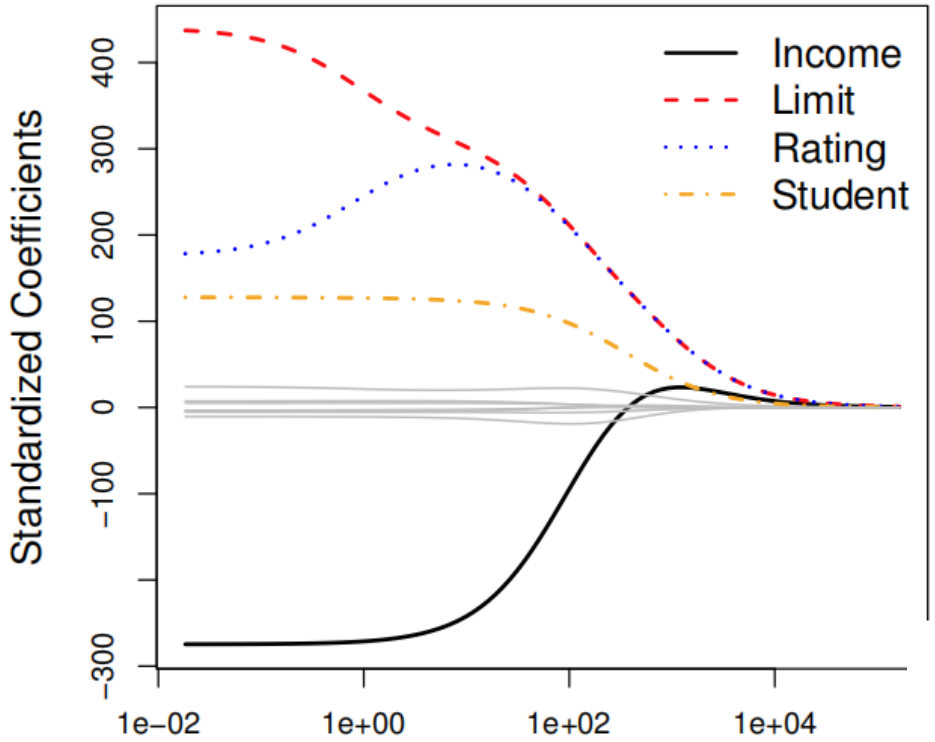
Ridge Regression (L2 Regularization)

- When $\boldsymbol{\lambda}$**(lambda)** $= \mathbf{0}$, ridge regression will produce the least squares estimates.
- As $\boldsymbol{\lambda}$ **increases**, the slope of the regression line is reduced and becomes more horizontal; the model becomes less sensitive to the variations of the independent variable
- As $\boldsymbol{\lambda} \rightarrow \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.



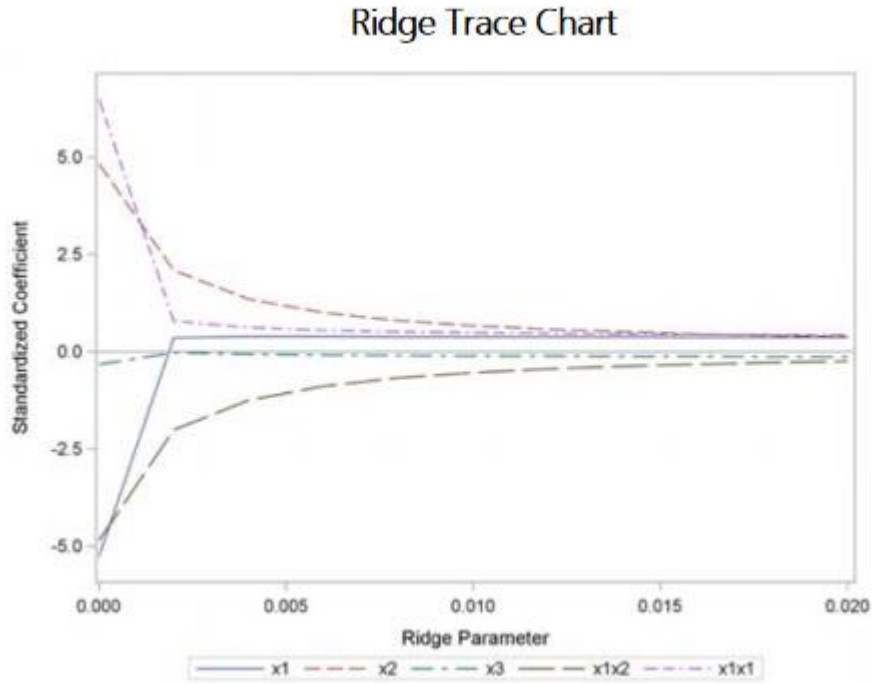
Ridge Regression (Ridge Trace)

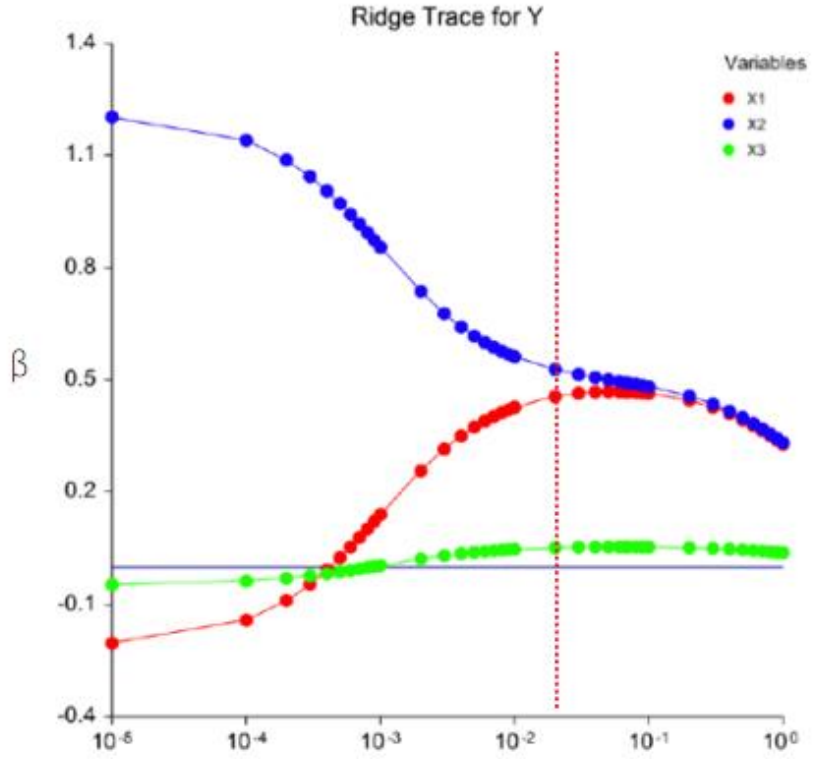
- Unlike least squares, which generate only one set of coefficient estimates, ridge regression will produce a different set of coefficient estimates, $\hat{\beta}_\lambda^R$ for each value of λ. [**Ridge Trace**]
- A ridge trace is a plot that shows the ridge regression coefficients as a function of λ.



Ridge Regression (Ridge Trace)

- When viewing the ridge trace we are looking for the λ that **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.

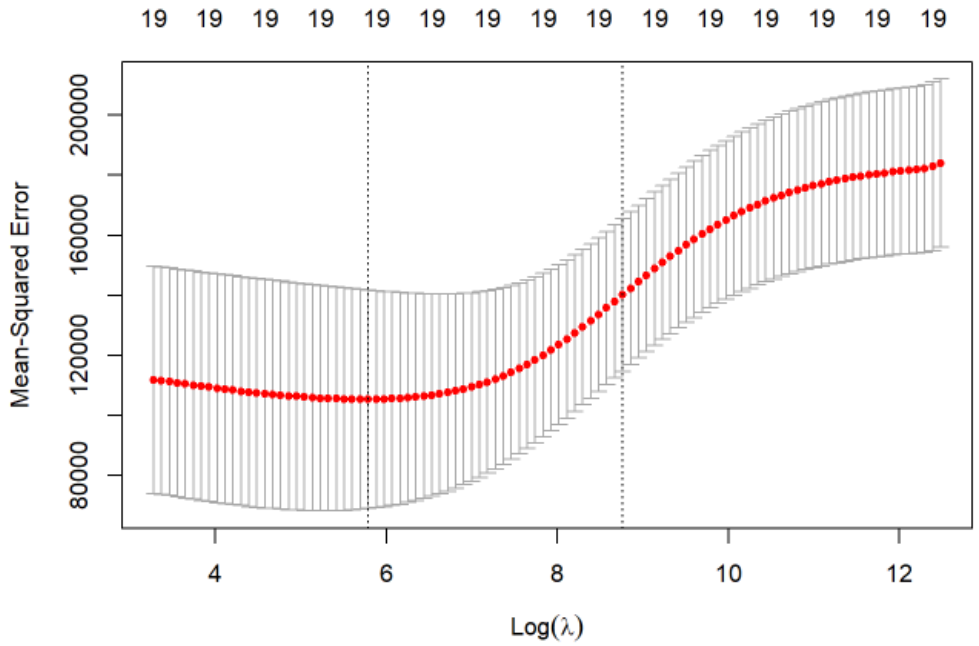




Ridge Regression (Ridge Trace)

- Alternatively, there are procedures in R which automatically selects the lowest value for λ .

```
set.seed(1)
cv.out <- cv.glmnet(x[train, ], y[train], alpha = 0)
plot(cv.out)
```




```
bestlam <- cv.out$lambda.min
bestlam
```

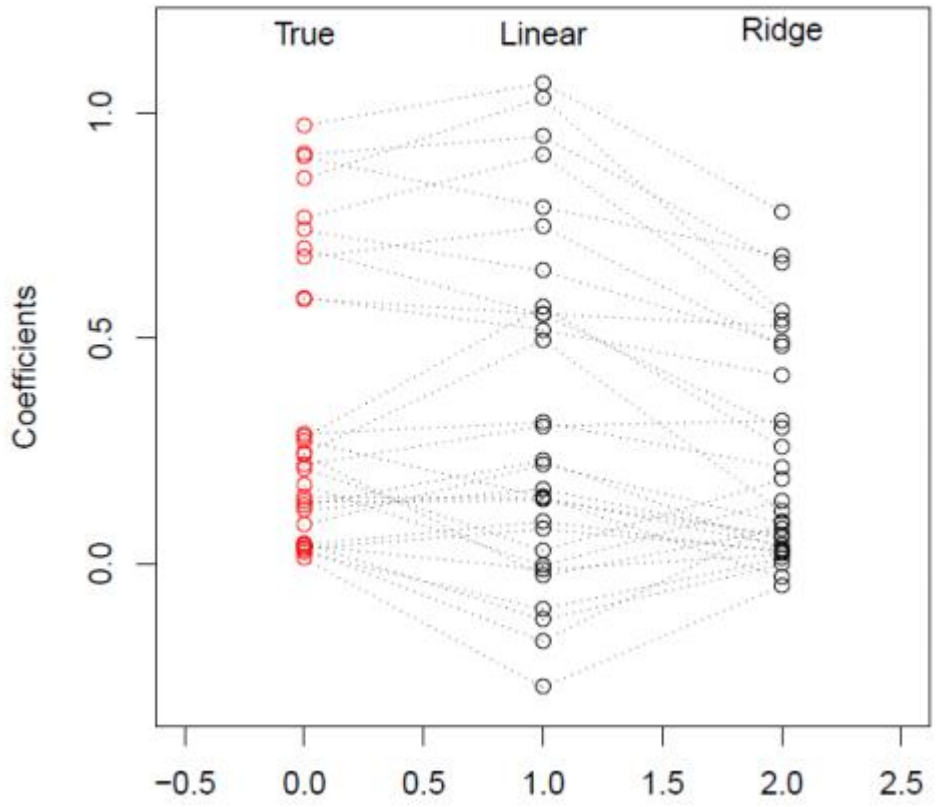
```
## [1] 326.0828
```

Ridge Regression (Scaling of Predictor)

- Here is a visual representation of the ridge coefficients for λ versus a linear regression

- The size of the coefficients (penalized) has decreased through our shrinking function.
- It is also important to point out that in ridge regression we usually leave the intercept unpenalized 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.**

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$



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**, i.e. will include all p predictors in the final model.
- 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 Regression (L1 Regularization)

- Lasso regression is similar to ridge regression
- It works by introducing a bias term but instead of squaring the slope, the **absolute value of the slope** is added as a penalty term

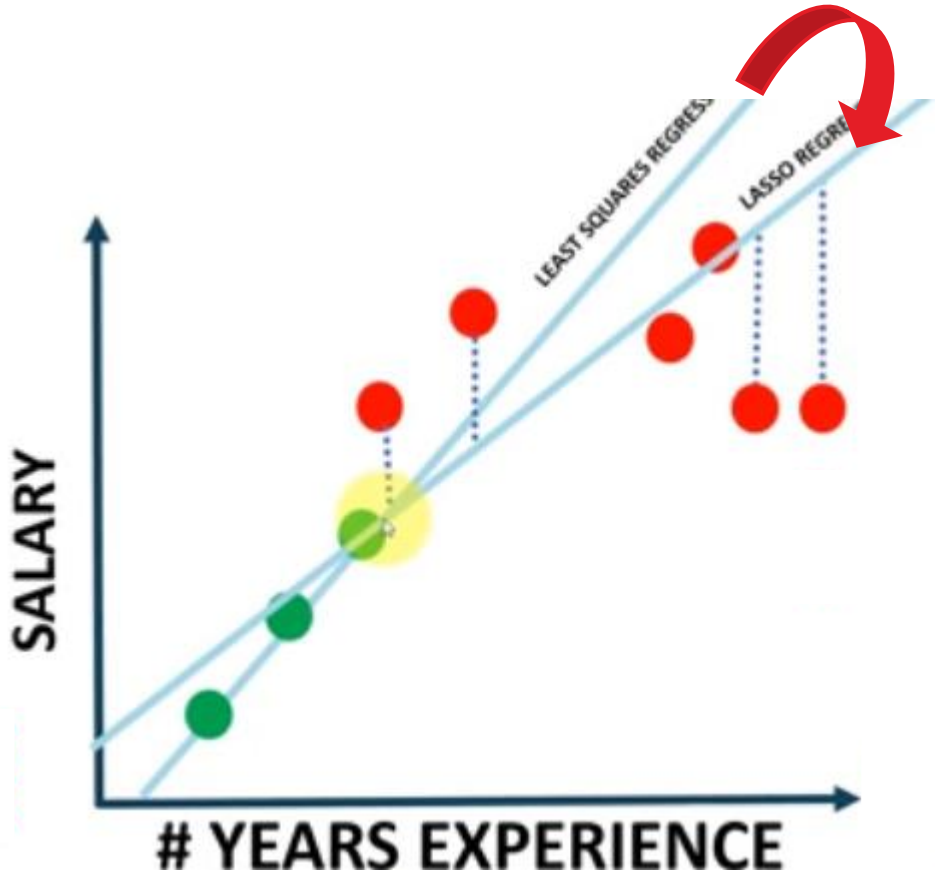
Least Square Regression:

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

Ridge Regression

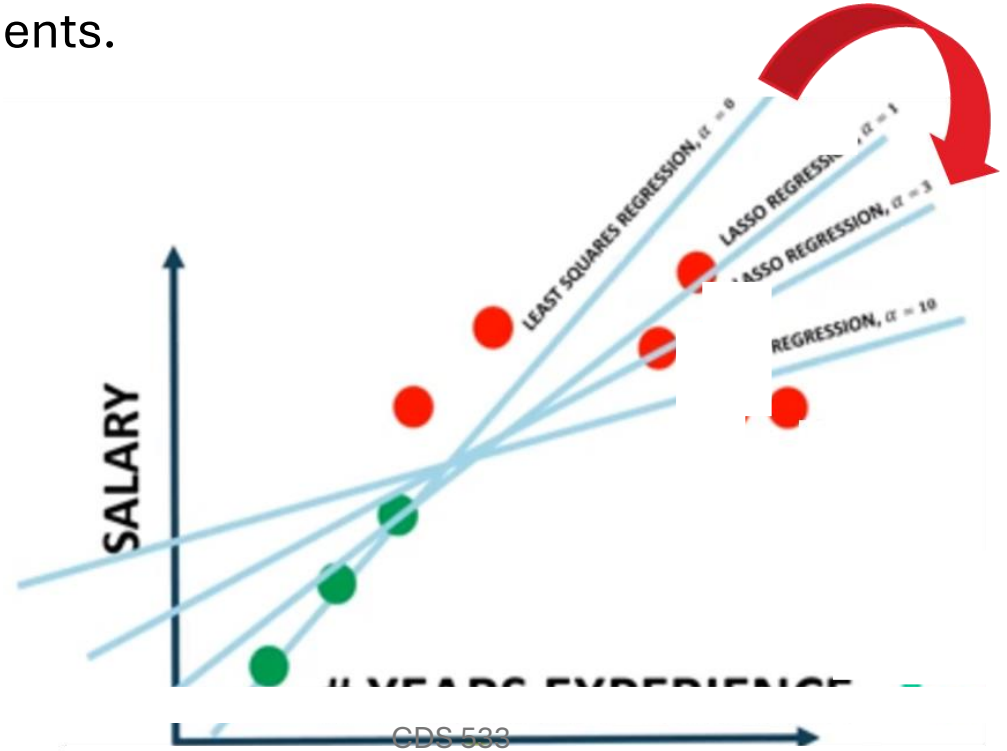
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

SHRINKAGE PENALTY TERM



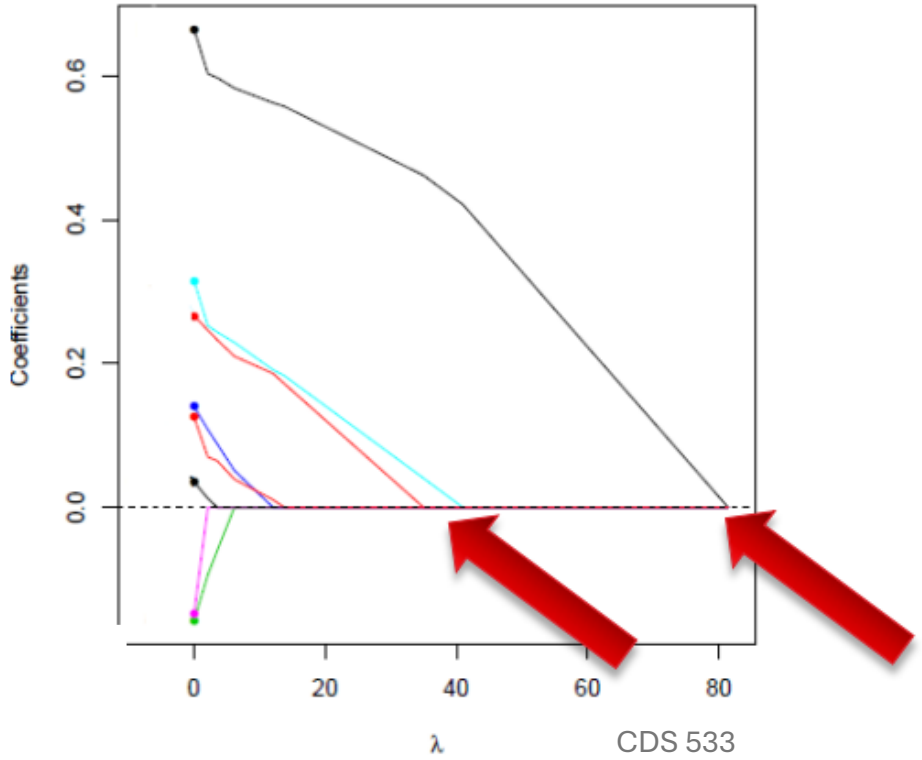
Lasso Regression (L1 Regularization)

- In lasso, the L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- **As λ increases**, the slope of the regression line is reduced and becomes more horizontal.
- **As λ increase**, the model becomes less sensitive to the variations of the independent variable.
- For λ in between these two extremes, we are balancing 2 ideas: fitting a linear model of y on X, and shrinking the coefficients.
-



Lasso Regression (L1 Regularization)

- The nature of the lasso penalty 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.
  - As λ increases, more coefficients are set to zero (less variable selected), and among non-zero coefficients, more shrinkage is employed.



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

Constrained Form

- It can be helpful to think about our penalty L1 and L2 parameters in the following form:

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

$$\hat{\beta}^{\text{ridge}} = \text{argmin } \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \quad \ell_2 \qquad \|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j{}^2}.$$

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

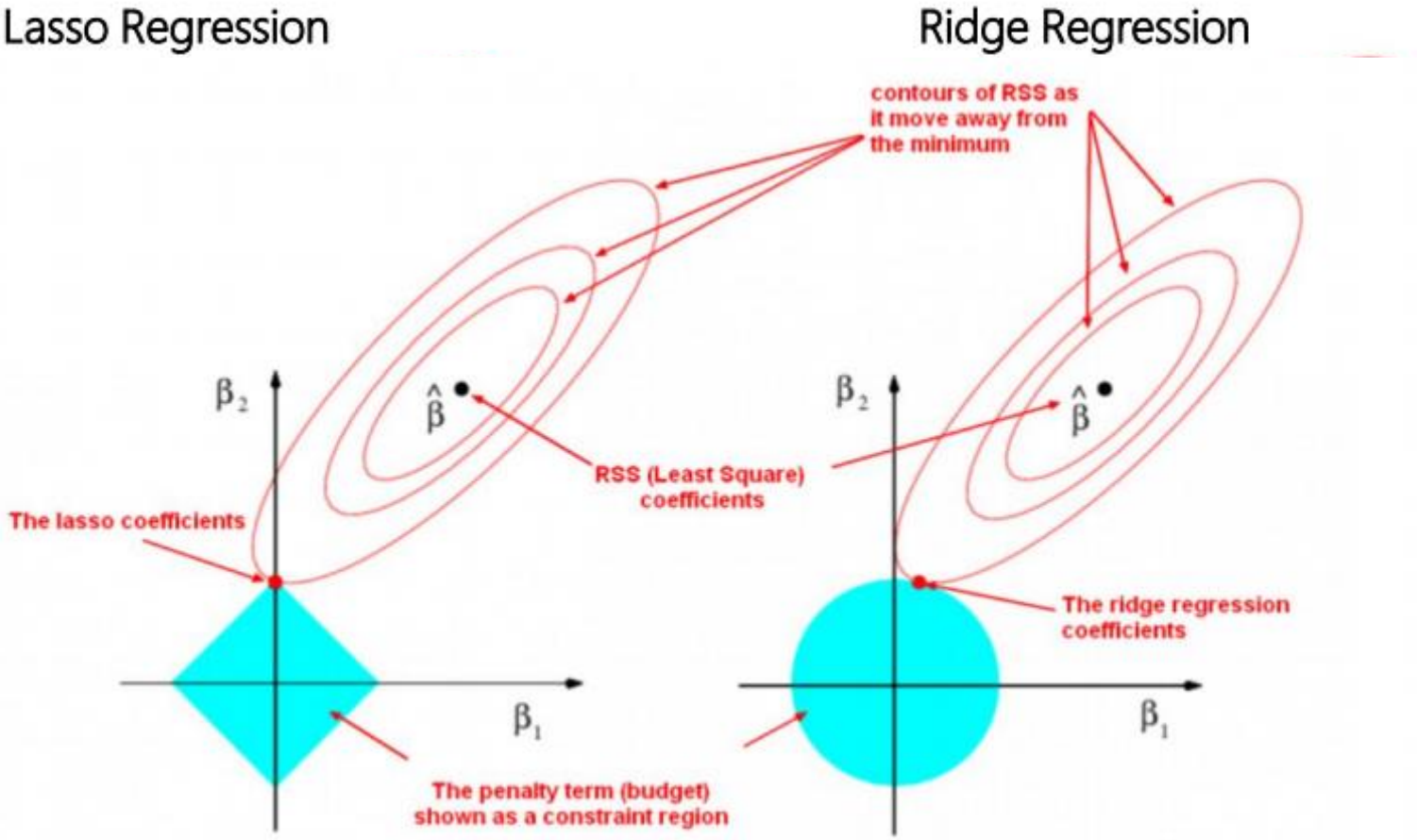
$$\hat{\beta}^{\text{lasso}} = \text{argmin } \|y - X\beta|_2^2 \quad \text{subject to } \|\beta\|_1 \leq t$$

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

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

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

Constrained Form



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

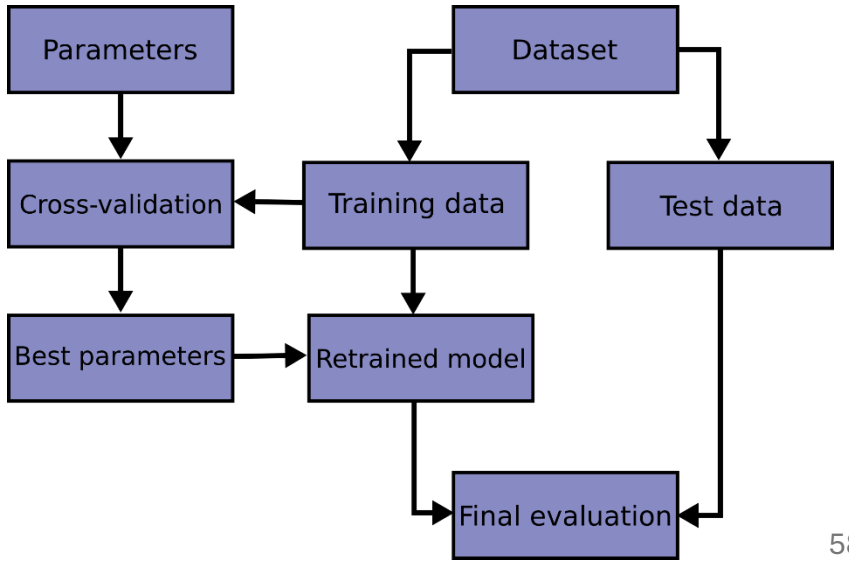
The contour lines are the least squares error function. The blue circle is the constraint region for the lasso regression.

Conclusion

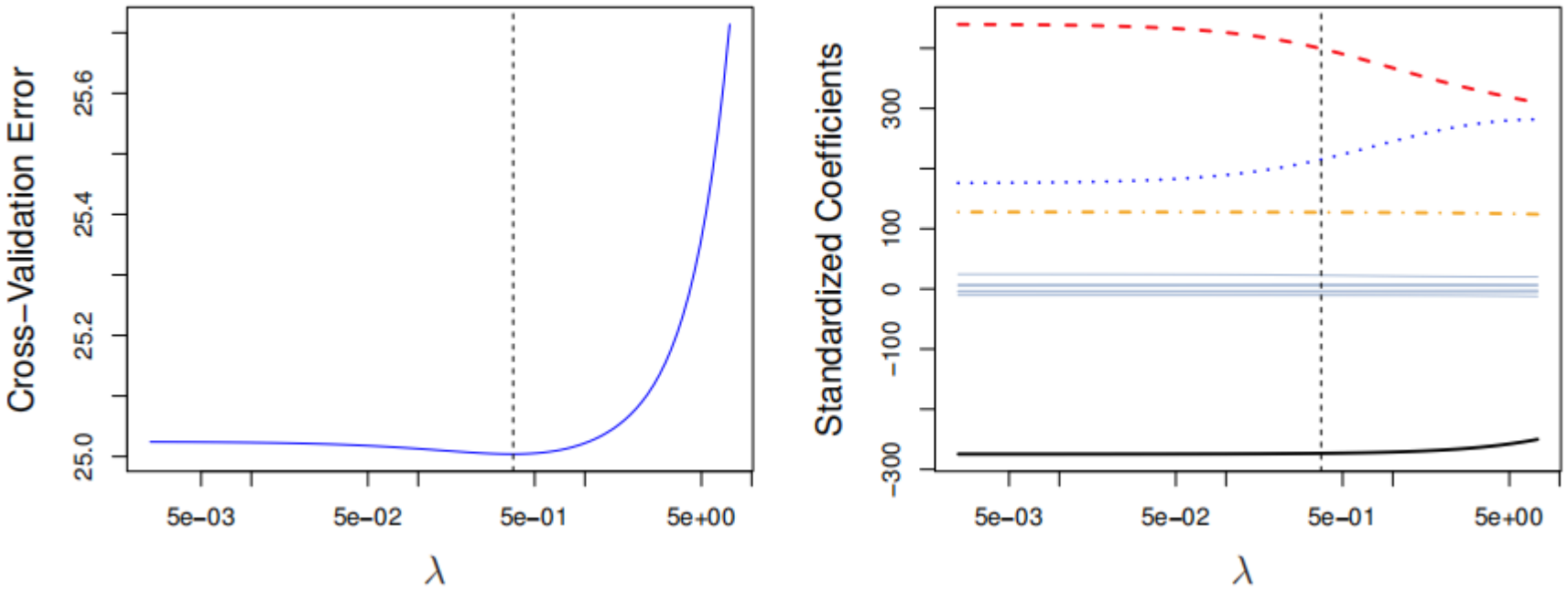
- In general, one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors.
- However, the number of predictors that is related to the response is never known a **priori** for real data sets.
- A technique such as cross-validation can be used in order to determine which approach is better on a particular data set.

Conclusion

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration is best.
- That is, we require a method selecting a value for the tuning parameter λ or equivalently, the value of the constraint t.
- Cross-validation provides a simple way to tackle this problem. We choose a grid of λ values, and compute the cross-validation error rate for each value of λ .
- We then select the tuning parameter value for which the cross-validation error is smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

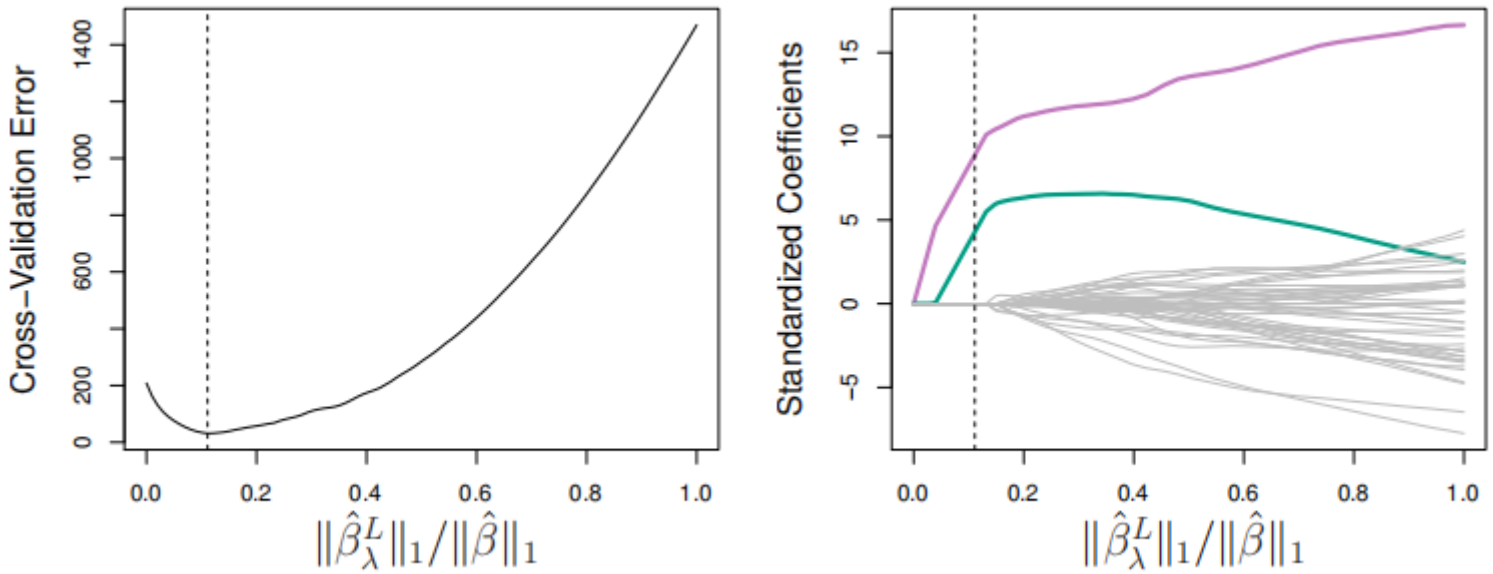


Credit dataset



- **Left:** Cross-validation errors that result from applying ridge regression to the Credit data set with various values of λ.
- **Right:** The coefficient estimates as a function of λ. The vertical dashed lines indicate the value of λ selected by cross-validation.

Credit dataset



- **Left:** Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set
- **Right:** The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.

Elastic Net Regression

- **Flexibility between ridge and lasso regression.**
- The elastic net forms a hybrid of the ridge and lasso 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 Regression

- **Flexibility between ridge and lasso regression.**
- Ridge, Lasso, and Elastic Net are all part of the same family with the penalty term of:

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

- 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

Regularization Part 1: Ridge (L2) Regression
Regularization Part 2: Lasso (L1) Regression
Regularization Part 3: Elastic Net Regression
Ridge, Lasso and Elastic-Net Regression in R

Lab Time!