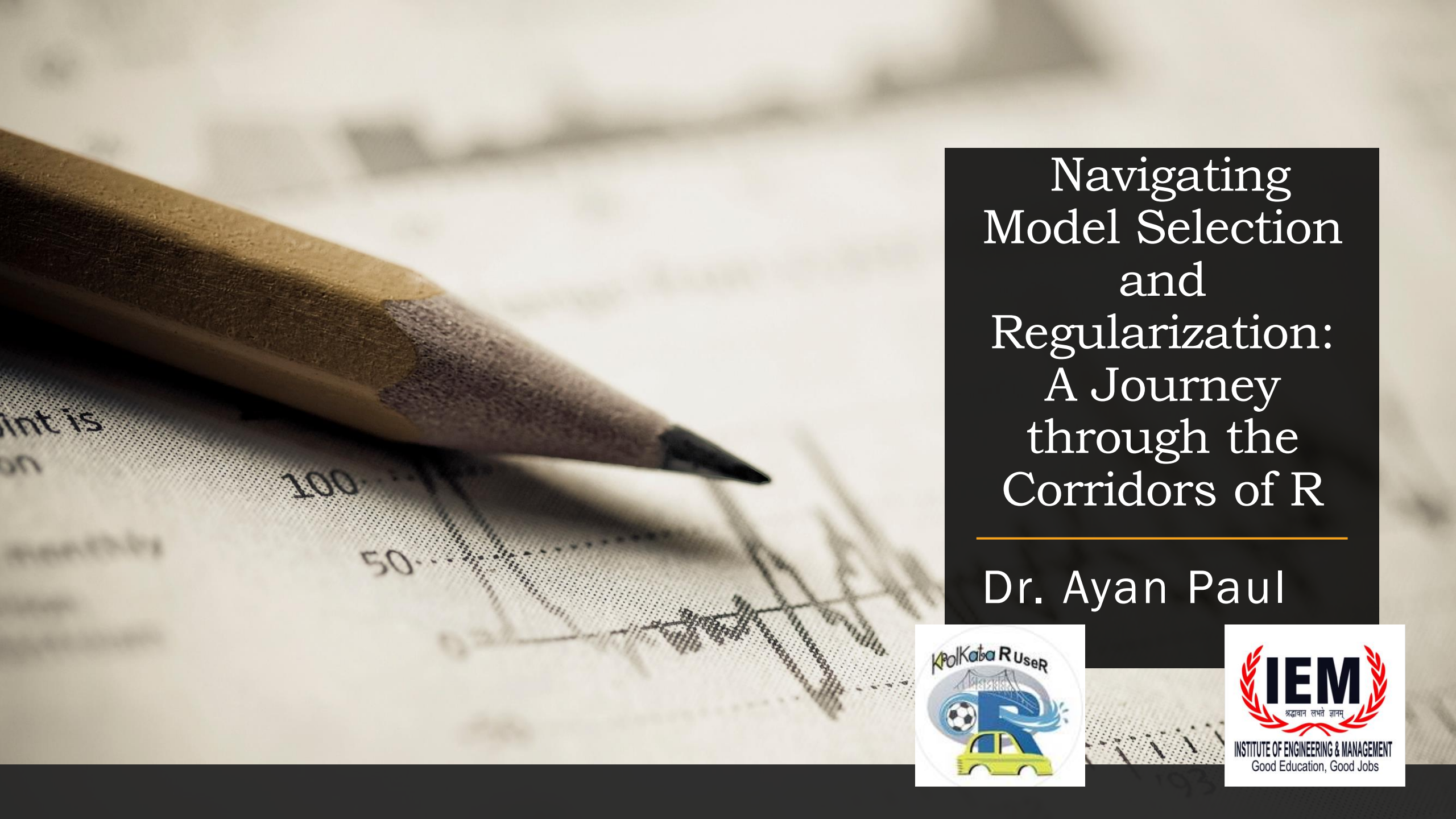


Navigating Model Selection and Regularization: A Journey through the Corridors of R

Dr. Ayan Paul

KolKata RUseR

IEM
श्रद्धावान लभते ज्ञानम्
INSTITUTE OF ENGINEERING & MANAGEMENT
Good Education, Good Jobs

Fundamentals of Data Science

Data

Training data | Testing data

Fit different models on Training data & choose any one

Datasets

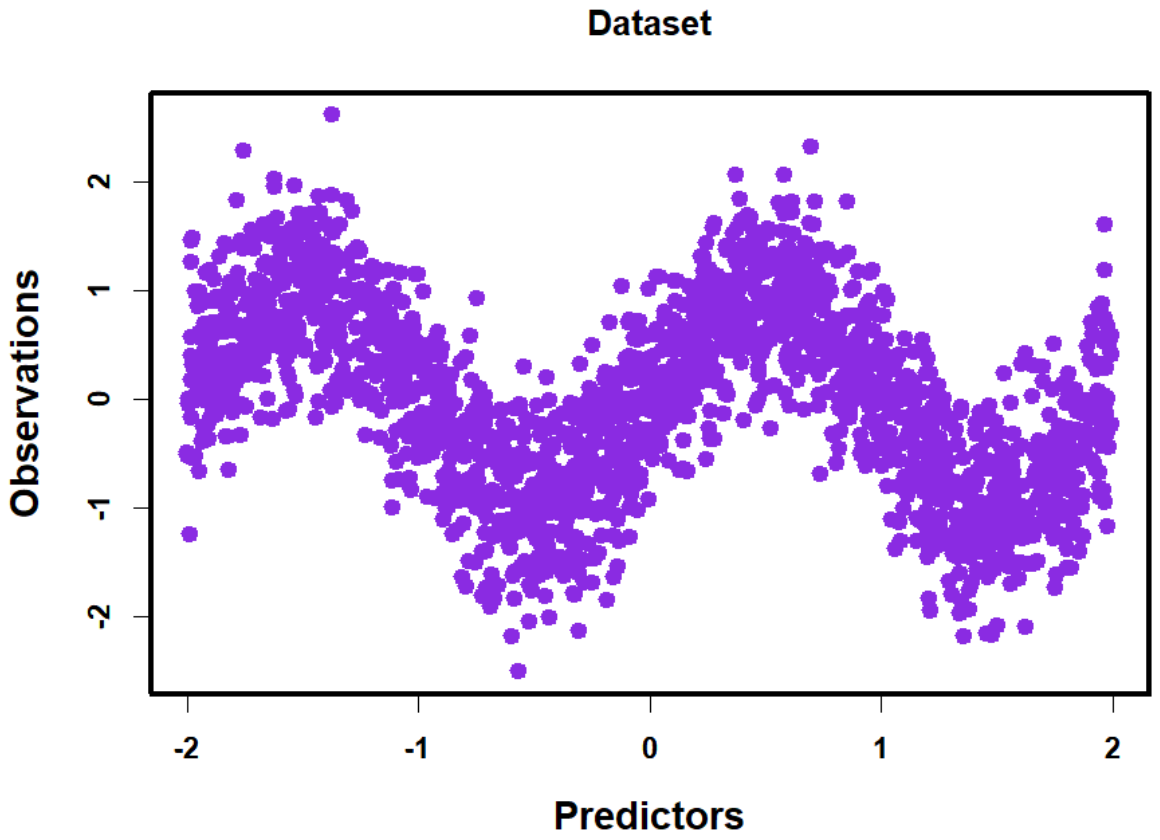
Fit the models by assuming that the **prediction accuracy** for both the train and test data are **almost identical**.

Apply chosen model to predict the test data

Simulation of the dataset

```
lines = par(lwd = 3)
set.seed(123)

# Simulate data
n = 2000
x = sort(runif(n, -2, 2))
f_true = function(x) {sin(pi * x)}
sigma2 = 0.25  # Variance of irreducible
error
y = f_true(x) + rnorm(n, mean = 0, sd =
sqrt(sigma2))
plot(x, y, pch = 19, xlab = "Predictors",
ylab = "Observations",
     cex.lab = 1.3, font.lab = 2, col =
"blueviolet", main = "Dataset")
axis(1, font = 2); axis(2, font = 2)
```



Partitioning of the dataset

```
# Split into training and test sets
train_id = sample(1:n, 70)
test_id = setdiff(1:n, train_id)

x_train = x[train_id]
y_train = y[train_id]
x_test = x[test_id]
y_test = y[test_id]
```

Key Questions

Is there a statistically significant disparity in **predictive accuracy** between the training and testing datasets?

If present, **do they propagate** any error within the model fitting process?

Are they **influenced by the degree of flexibility** in the model parameters?

Measurement of Accuracy

Mean Squared Error (MSE)

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f(x_i)}\right)^2$$

n = sample Size
y_i = True Response Values
$\widehat{f(x_i)}$ = Prediction of i -th observations.

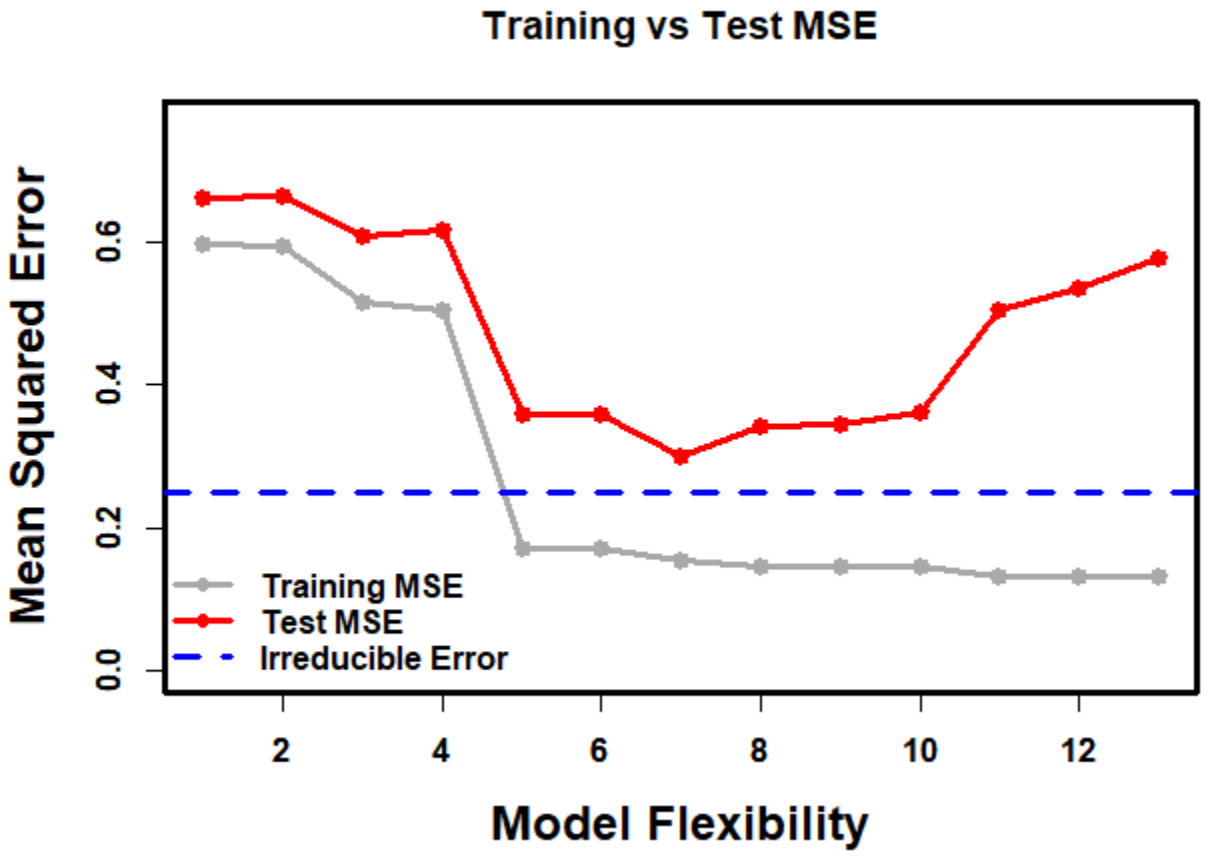
Evaluation of Test and Train MSE of the data

```
max_degree <- 13
train_mse <- numeric(max_degree)
test_mse <- numeric(max_degree)
for (i in 1:max_degree) {
          # Create polynomial regression model
  form <- as.formula(paste("y_train ~ poly(x_train, ", d, ",
raw=TRUE)", sep = ""))
model <- lm(form)
# Predictions
y_pred_train <- predict(model)
y_pred_test <- predict(model, newdata = data.frame(x_train =
x_test))
train_mse[i] <- mean((y_pred_train - y_train)^2)
test_mse[i] <- mean((y_pred_test - y_test)^2)
}
```

Understanding between Train and Test MSE

Key Observations

- With increasing model flexibility, the training MSE exhibits a consistently **declining trajectory**
- However, **this trend may not necessarily hold** for the test mean squared error (MSE).
- Beyond a certain threshold of model flexibility, the test MSE begins to rise, exhibiting a characteristic **U-shaped trend**.
- As model flexibility increases beyond a threshold, a **substantial divergence** between the training and test MSE becomes evident.

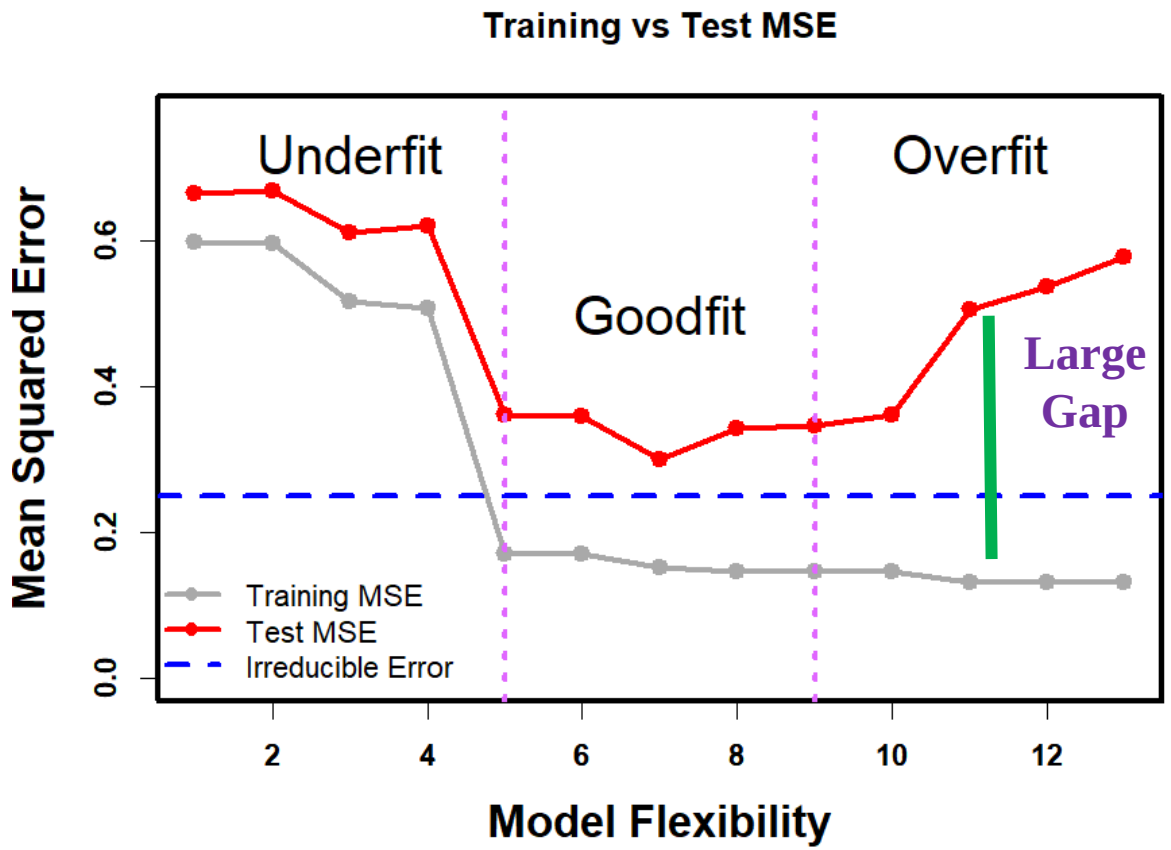


A First Step towards Model Selection

Fitting type	Train MSE	Test MSE
Under fit	High	High
Good fit	Low (not close to zero)	Low
Over fit	Low	High

Take Home Message

- Low model flexibility is not good.
- Again, High Model flexibility is also not good.
- Moderate level of flexibility is required for the model selection process.



Outline of Model Selection

- **Structure of General Linear Regression Model**

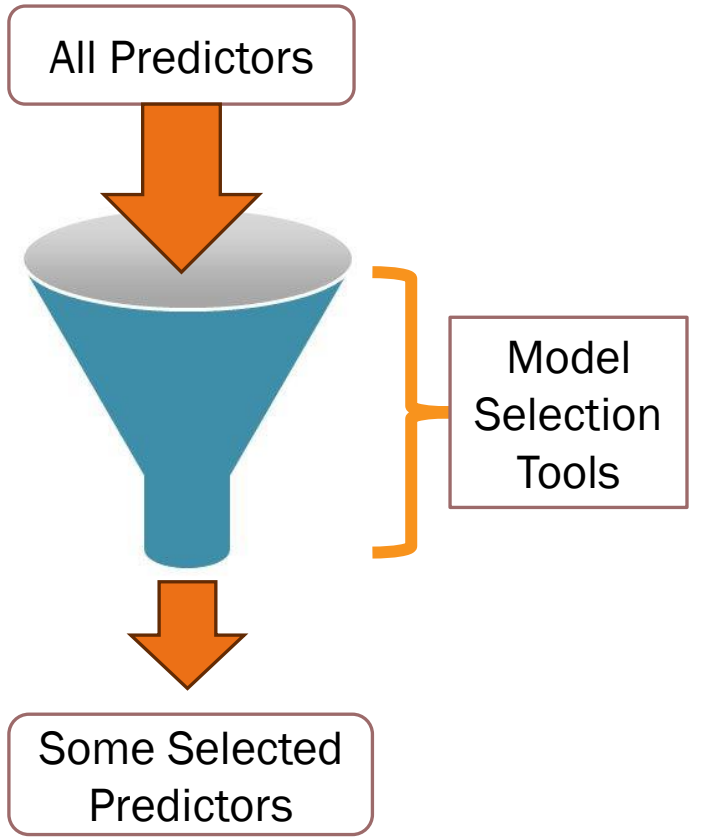
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \cdots + \beta_p X_p + \epsilon$$

where Y is the response variable and $X = (X_1, X_2, X_3, \ldots, X_p)$ be the set of predictors or explanatory variable.

- **R code to execute multiple linear regression model**

```
lm(response ~ predictor (1) + predictor (2) + predictor (3), data = datafile)
```

- **Are all predictors necessary to explain the linear systems?.....**
- **Begins the brainchild of model selection problem.....**



Need of Model Selection

- It is often the case that some or many of the variables used in a multiple regression model **are in fact not associated** with the response.
- **Including such irrelevant variables** leads to **unnecessary complexity** in the resulting model.
- **Model selection is mainly required for two agenda**
- **Prediction accuracy:** This strategy deals with the **number of observations (n)** and **number of predictors (p)**

If $n > p$ then it is ok. **We can go with all the predictors**.
If $n < p$ then **selection of predictor is needed** since the variance of the $\hat{\beta}$ will increase and tends to infinite.

- **Model Interpretability:** Since consideration of **unnecessary predictors will produce poor estimates** of the model parameter so, it becomes hard to interpret the underlying system.

Model Selection Process

The Model selection can be performed by the following three ways

- **Subset Selection:** The process involves in selecting the subset of predictors among the set of p many predictors. We then use linear regression model on that subset to find the estimates.

  Statistical Process: Stepwise Forward Regression, Stepwise Backward Regression

- **Shrinkage Method:** This approach involves fitting a model involving all p predictors. But, some of the estimated coefficients are converging to zero. Consequently the variance of the model parameter estimates becomes reduced.

  Statistical Process: Lasso, Ridge Regression

- **Dimension Reduction:** This approach involves projecting the p predictors into a M −dimensional subspace, where $M < p$.

  Statistical Process: Principal Component Analysis (PCA)

Live Demonstration

Response Variable

The data is related to some physio-chemical parameters of a molluscan species

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	pop	ST	SpH	Ssal	SORP	WT	WpH	DO	Wsal	pneu	ACO2	AT	Sand	Silt	Clay	SM	BD	SP	EC	SOC	TOM	Sphos	Sacid	STN	TDS
2	8	32	8.4	7.3	-138	31.5	8.5	6.2	19.6	235	628	32.6	70.3	18.3	11.47	162.73	0.97	56.51	6.43	4.64	7.99	11.88	1.01	0.05	19210
3	9	31.9	8.6	7.2	-140	31.4	8.4	6.2	19.5	238	560	31.9	66.84	22.06	11.1	92.67	0.82	55.37	6.31	4.53	7.8	10.9	1.02	0.05	18860
4	8	33.7	7.8	9.2	-133	31.2	8.2	6.4	20.5	247	519	31.6	64.73	23.17	12.1	88.89	0.91	57.01	6.11	4.87	8.48	18.43	0.34	0.09	18560
5	8	33.6	7.9	9	-128	30	8.3	6.6	20.2	231	548	31.8	68.73	18.97	12.3	120.65	0.89	53.07	6.84	4.33	7.67	16.89	0.78	0.06	18930
6	9	33.2	7.8	12.9	-122	31.9	8.4	6.7	20	241	586	32	67.73	20.37	11.9	106.47	0.86	54.65	6.78	4.49	7.86	21.32	0.84	0.12	18490
7	10	32	8.5	7.3	-144	31.7	8.6	6.3	19.6	240	740	32.7	55.23	33.3	11.47	70.9	0.7	54.55	6.59	4.87	8.39	13.69	0.09	0.06	18290
8	5	33.7	8.1	13.1	-68	33.2	8.6	7.8	20.5	300	626	32.2	56.45	29.55	14	40.54	0.81	60.49	6.53	5.19	8.94	46.93	1.3	1.8	18830
9	8	31.8	8.3	9.3	-45	32	8.5	6.7	20.3	275	389	31.7	59.88	29.55	10.57	37.34	0.98	58.84	7	4.73	8.15	11.29	0.14	0.04	18880
10	9	31.6	7.9	12.6	-69	31.8	8.4	7.6	20.4	268	512	31.9	58.8	30.8	10.6	28.98	0.93	58.7	5.64	4.56	7.86	26.68	0.18	0.78	18120
11	5	31.5	7.7	12.3	-75	31.8	8.6	7.5	20.2	281	588	32	46.45	33.3	20.29	37.78	0.76	53.74	6.83	4.73	8.15	9.66	1.1	1.4	17910
12	4	30.3	8.3	23.4	-114	30.7	8.8	8	21	342	864	30.7	41.65	36.5	21.85	32.29	0.98	56.64	5.69	4.9	8.44	13.33	0.29	0.08	19180
13	6	29.4	8.2	12.1	-102	30.2	8.3	7.6	22.6	382	313	30.3	43.25	38.65	18.1	33.88	0.97	48.68	7.11	5.3	9.15	29.33	0.22	0.01	22560
14	4	30.3	8.1	23.1	-103	30.4	8.6	7.9	19.7	292	268	30.6	43.25	37.4	19.35	42.56	0.97	40.5	7.16	4.43	7.63	39.95	0.11	0	16970
15	5	30.2	7.8	21.6	-106	30.5	8.7	7.7	21	289	145	30.5	44.64	34.74	20.62	38.66	0.88	51.01	5.99	3.82	11.75	31.56	0.49	0.02	17060
16	3	29.3	8	21.4	-99	30.6	8.5	7.1	22.4	311	264	30.2	32.9	41.5	25.6	33.26	1	53.01	6.06	4.69	8.08	15.7	0.15	0.01	19600
17	0	25.7	8.1	11.3	73	27.6	8.4	9.9	23.7	114	352	31.3	3.32	43.21	53.47	48.92	1.22	61.13	7.89	2.67	4.6	9.908	0.89	0.073	16980
18	2	24.9	8.1	11.6	148	26.5	8.4	8.3	25.5	52	242	28.6	3.76	61.02	35.22	49.72	1.22	59.08	6.46	0.82	1.41	0.028	0.64	0.085	17210
19	0	24.2	7.9	11.4	63	26.1	8.2	8.5	26	173	306	27.3	3.35	67.12	29.53	46.89	1.18	60.49	6.72	3.23	5.57	19.293	0.79	0.061	18280
20	1	27.7	7.8	13.5	79	26.6	8.2	9.6	24	106	348	30.9	4.01	58.69	37.3	51.93	1.12	48.61	6.87	2.89	4.98	18.081	0.81	0.089	17970
21	1	24.6	7.9	11.6	148	26.3	8.5	7.8	25.7	78	278	28.9	3.59	63.69	32.72	49.81	1.16	54.64	6.28	3.08	5.31	16.081	0.63	1.001	18670
22	2	27.7	8.1	13.5	79	27.6	8.4	9.9	23.9	114	352	31.9	3.32	43.21	53.47	44.07	1.22	58.84	7.13	2.67	4.6	9.908	0.89	0.073	24060
23	2	29.7	8	10.3	103	29.7	8.6	8.7	26.3	172	145	30	5.7	58.59	35.71	44.98	1.21	56.09	6.49	2.82	4.86	0.028	0.64	0.072	23500
24	4	25.7	7.9	13.3	67	26.7	8.3	9.7	23.8	102	344	31.9	4.36	56.96	38.68	44.62	1.07	54.98	6.74	3.08	5.37	11.083	0.39	0.094	22700
25	5	25.6	7.7	11.4	62	26.6	8.2	9.5	23.7	121	342	31.6	5.33	55.96	38.71	44.38	1.29	58.72	6.36	3.28	5.71	12.09	0.58	0.098	23210
26	3	29.6	7.8	9.6	96	29.1	8.6	8.5	26	162	92	29.5	5.16	57.96	36.88	45.16	1.13	57.38	6.02	3.42	5.96	14.087	0.62	0.087	21960
27	4	32.2	8	13.1	-66	32	8.8	4.1	31.4	154	376	31.9	48.05	33.3	18.65	30.46	0.76	55	11.23	4.81	8.29	8.17	15	3	30680

Coefficients:

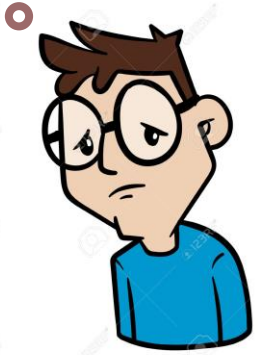
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.174e+02	8.474e+02	0.257	0.8010	
ST	-4.608e-01	3.557e-01	-1.296	0.2147	
SpH	-1.785e+00	1.696e+00	-1.053	0.3092	
Ssal	-9.298e-02	8.154e-02	-1.140	0.2720	
SORP	-1.213e-02	9.874e-03	-1.229	0.2382	
WT	2.320e-01	4.845e-01	0.479	0.6389	
WpH	2.144e+00	2.739e+00	0.783	0.4460	
DO	7.539e-01	4.781e-01	1.577	0.1357	
Wsal	7.177e-02	2.159e-01	0.332	0.7442	
pneu	-1.352e-02	7.603e-03	-1.779	0.0956	.
ACO2	-2.132e-03	2.667e-03	-0.799	0.4366	
AT	4.017e-01	2.506e-01	1.603	0.1298	
Sand	-2.169e+00	8.469e+00	-0.256	0.8013	
Silt	-2.289e+00	8.489e+00	-0.270	0.7911	
Clay	-2.406e+00	8.493e+00	-0.283	0.7808	
SM	-1.374e-02	1.646e-02	-0.835	0.4168	
BD	2.372e-04	2.721e-04	0.872	0.3970	
SP	7.428e-02	4.186e-02	1.774	0.0963	.
EC	-4.029e-01	3.498e-01	-1.152	0.2674	
SOC	7.328e-01	8.463e-01	0.866	0.4002	
TOM	-3.745e-01	4.201e-01	-0.892	0.3867	
Sphos	-1.051e-02	4.115e-02	-0.255	0.8019	
Sacid	4.205e-02	2.641e-01	0.159	0.8756	
STN	-1.128e+00	9.437e-01	-1.196	0.2504	
TDS	1.190e-04	7.466e-05	1.594	0.1317	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Coefficient estimates of Linear Regression

No parameter is significant !!

p-value



Variance of the model parameter estimates associated with the linear regression

(Intercept)	ST	SpH	Ssal	SORP	WT	WpH	DO	Wsal	pneu	ACO2
7.180466e+05	1.264947e-01	2.875528e+00	6.648490e-03	9.748887e-05	2.347274e-01	7.503444e+00	2.285743e-01	4.660593e-02	5.780396e-05	7.114810e-06
AT	Sand	Silt	Clay	SM	BD	SP	EC	SOC	TOM	Sphos
6.278564e-02	7.171615e+01	7.206399e+01	7.212447e+01	2.707745e-04	7.403162e-08	1.752542e-03	1.223534e-01	7.161612e-01	1.764775e-01	1.693498e-03
Sacid	STN	TDS								
6.976462e-02	8.905055e-01	5.573859e-09								

Summary of the Best Subset Selection Process

	ST	SpH	Ssal	SORP	WT	WpH	DO	Wsal	pneu	ACO2	AT	Sand	Silt	Clay	SM	BD	SP	EC	SOC	TOM	Sphos	Sacid	STN	TDS
1 (1)														*										
2 (1)													*								*			
3 (1)								*				*					*							
4 (1)								*				*					*				*			
5 (1)								*				*					*				*			
6 (1)							*					*	*	*							*			
7 (1)							*		*			*	*	*							*			*	
8 (1)							*		*			*	*	*			*				*			*	
9 (1)							*		*			*	*	*			*				*			*	
10 (1)						*			*			*	*	*			*				*			*	*
11 (1)					*	*			*			*	*	*			*				*			*	*
12 (1)					*	*			*			*	*	*			*	*			*			*	*
13 (1)					*	*			*			*	*	*			*	*			*			*	*
14 (1)	*	*	*	*			*	*				*	*	*			*				*			*	*
15 (1)	*	*	*	*			*	*				*	*	*			*	*			*			*	*
16 (1)	*	*	*	*			*	*				*	*	*			*	*	*		*			*	*
17 (1)	*	*	*	*	*		*					*	*				*	*	*		*			*	*
18 (1)	*	*	*	*			*	*			*	*					*	*	*	*	*			*	*
19 (1)	*	*	*	*			*	*			*	*					*	*	*	*	*			*	*
20 (1)	*	*	*	*			*	*	*		*	*					*	*	*	*	*			*	*
21 (1)	*	*	*	*	*		*	*	*		*	*					*	*	*	*	*			*	*
22 (1)	*	*	*	*	*		*	*	*		*	*					*	*	*	*	*			*	*
23 (1)	*	*	*	*	*		*	*	*		*	*	*				*	*	*	*	*			*	*
24 (1)	*	*	*	*	*		*	*	*		*	*	*				*	*	*	*	*	*		*	*

Subset Selection Process

The subset selection process includes three methods

(a) Best Subset Selection Method:

Algorithm 6.1 *Best Subset Selection*

1. Let $\mathcal{M}_0$ 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 $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Possible metrics to be used

RSS = Residual Sum of Square
C_p = Mallows C_p
AIC = Akaike Information Criterion
BIC = Bayesian Information Criterion

Variance of the model parameter estimates associated with the linear regression

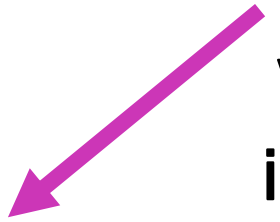
(Intercept)	ST	SpH	Ssal	SORP	WT	WpH	DO	Wsal	pneu	ACO2
7.180466e+05	1.264947e-01	2.875528e+00	6.648490e-03	9.748887e-05	2.347274e-01	7.503444e+00	2.285743e-01	4.660593e-02	5.780396e-05	7.114810e-06
AT	Sand	Silt	Clay	SM	BD	SP	EC	SOC	TOM	Sphos
6.278564e-02	7.171615e+01	7.206399e+01	7.212447e+01	2.707745e-04	7.403162e-08	1.752542e-03	1.223534e-01	7.161612e-01	1.764775e-01	1.693498e-03
Sacid	STN	TDS								
6.976462e-02	8.905055e-01	5.573859e-09								

One Predictor

(Intercept)	clay
0.4738261172	0.0007107512

Five Predictors

(Intercept)	Wsal	AT	Clay	SP	Sphos
6.0723897444	0.0035602127	0.0087187229	0.0004562867	0.0001462520	0.0005258906

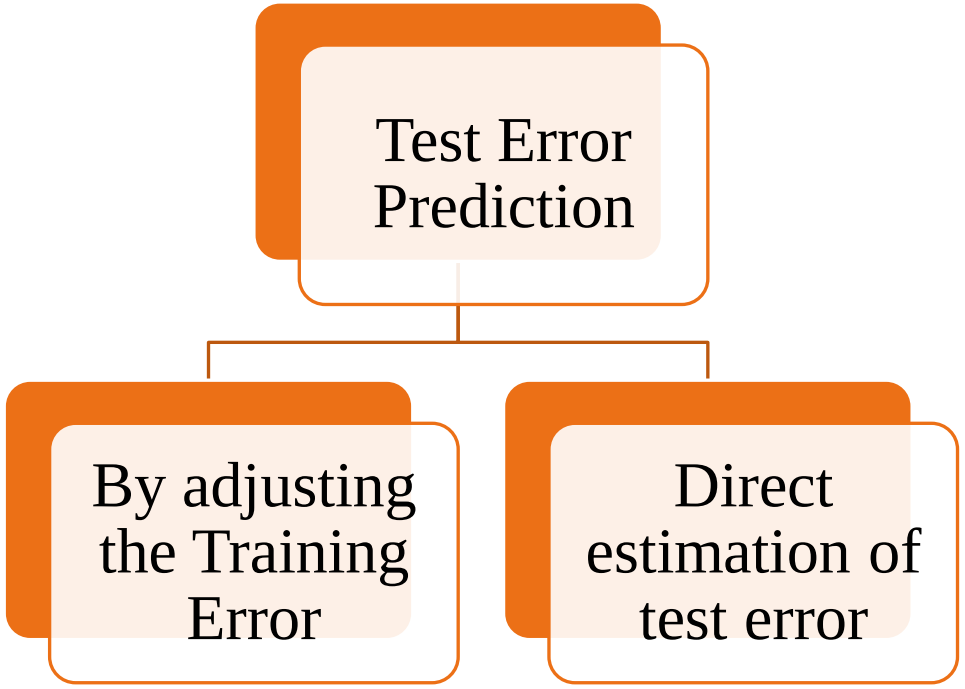


Variance increases

Ten Predictors

(Intercept)	SORP	DO	pneu	AT	Sand	Silt	SM	SP	STN	TDS
3.382968e+01	2.440032e-05	6.025978e-02	1.866907e-05	1.142701e-02	1.149958e-03	2.042187e-03	1.072752e-04	3.161316e-04	1.630536e-01	1.779977e-09

Optimal Model Selection



Training Error Adjustment

1. Mallow's $C_p = \frac{1}{n}(RSS + 2p\,\hat{\sigma}^2)$

2. AIC $= \frac{1}{n\hat{\sigma}^2}(RSS + 2p\,\hat{\sigma}^2)$

3. BIC $= \frac{1}{n}(RSS + \log(n)p\,\hat{\sigma}^2)$

Protocol

Lower the value, better will be the model

Variance of the model parameter estimates associated with the linear regression

(Intercept)	ST	SpH	Ssal	SORP	WT	WpH	DO	Wsal	pneu	ACO2
7.180466e+05	1.264947e-01	2.875528e+00	6.648490e-03	9.748887e-05	2.347274e-01	7.503444e+00	2.285743e-01	4.660593e-02	5.780396e-05	7.114810e-06
AT	Sand	Silt	Clay	SM	BD	SP	EC	SOC	TOM	Sphos
6.278564e-02	7.171615e+01	7.206399e+01	7.212447e+01	2.707745e-04	7.403162e-08	1.752542e-03	1.223534e-01	7.161612e-01	1.764775e-01	1.693498e-03
Sacid	STN	TDS								
6.976462e-02	8.905055e-01	5.573859e-09								

Best number (7) of Predictors

(Intercept)	DO	pneu	AT	Sand	Silt	SP	STN
2.977936e+01	4.669570e-02	1.140033e-05	1.056094e-02	8.251019e-04	1.761784e-03	1.613948e-04	1.143472e-01

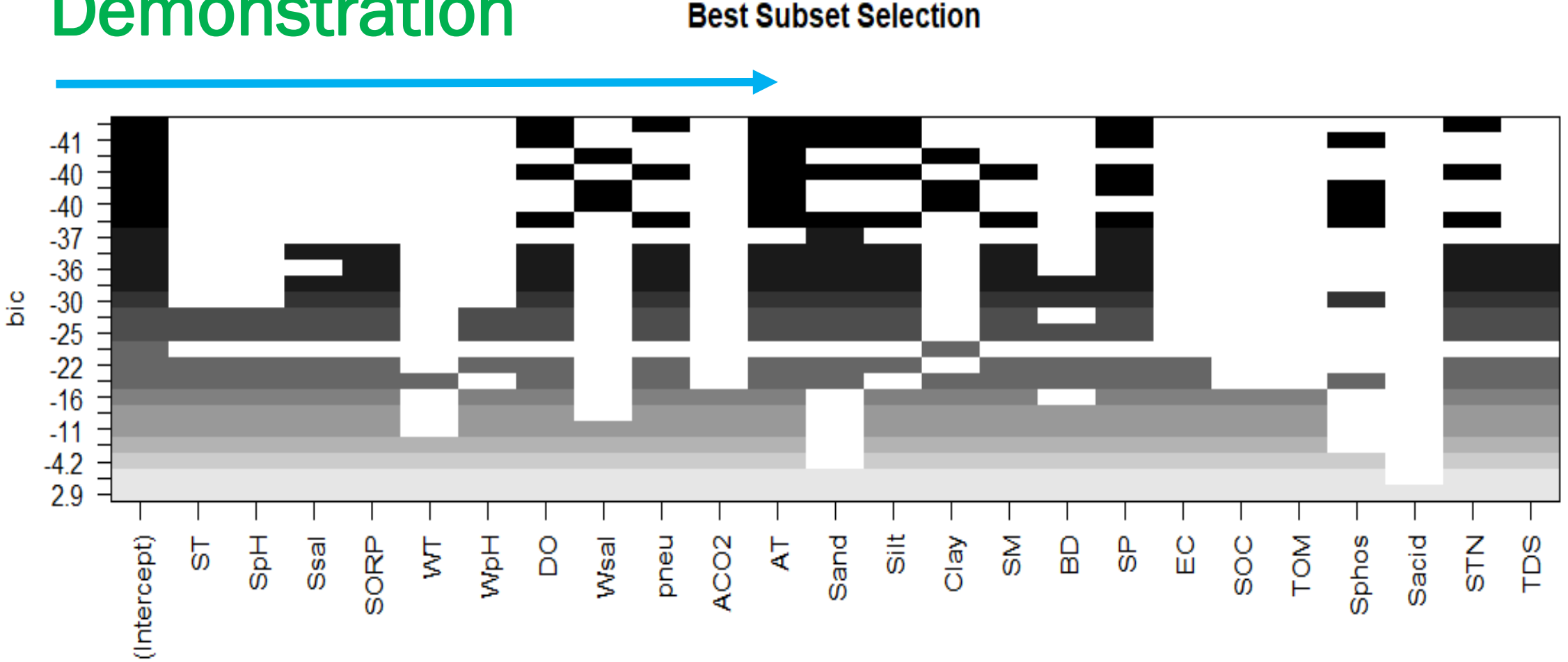
Ten Predictors

(Intercept)	SORP	DO	pneu	AT	Sand	Silt	SM	SP	STN	TDS
3.382968e+01	2.440032e-05	6.025978e-02	1.866907e-05	1.142701e-02	1.149958e-03	2.042187e-03	1.072752e-04	3.161316e-04	1.630536e-01	1.779977e-09

Graphical Demonstration

Protocol

Lower the value, better will be the model



(Intercept)	DO	pneu	AT	Sand	Silt	SP	STN
2.977936e+01	4.669570e-02	1.140033e-05	1.056094e-02	8.251019e-04	1.761784e-03	1.613948e-04	1.143472e-01

Stepwise Selection Process

(b) Stepwise Forward Selection Method:

Algorithm 6.2 *Forward Stepwise Selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Stepwise Selection Process (contd..)

(c) Stepwise Backward Regression Method:

Algorithm 6.3 *Backward Stepwise Selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Demonstration by R Software

Required Package for Executing Subset Selection Process: *leaps*

Required Command: *regsubsets*

Take Home Message

We should select the models with less number of predictors.

Reference Book

An Introduction to Statistical Learning

with Applications in R

Gareth James Daniela Witten Trevor Hastie
Robert Tibshirani

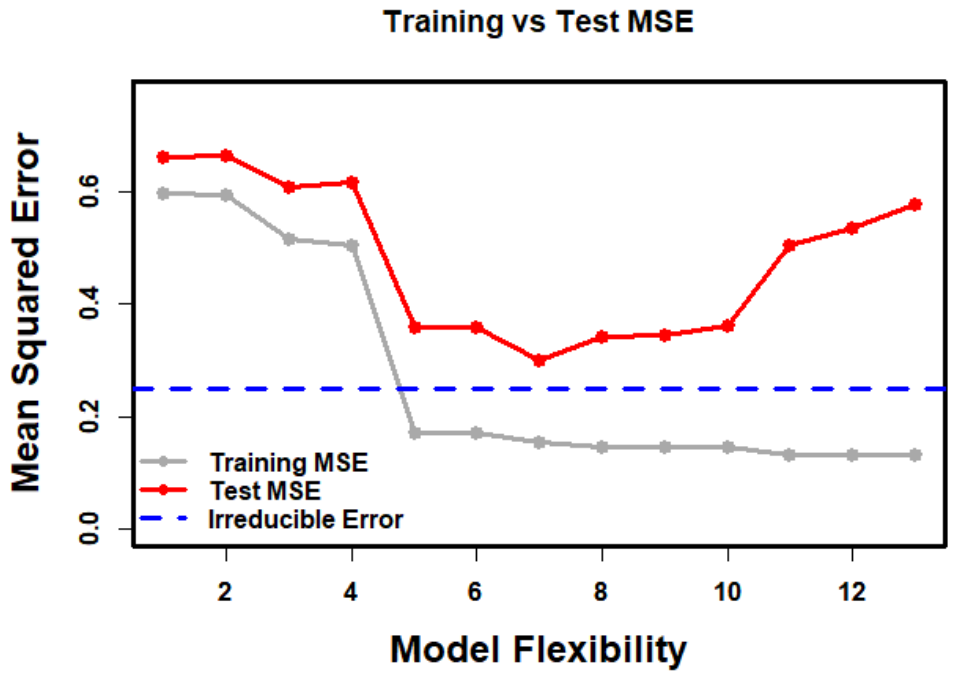
February 11, 2013

Thanks for listening

Test Mean Square Error (Test MSE)

Test MSE = Variance of test observations + Bias of test observations + Irreducible Error

- Variance of test observations > 0
- Bias of test observations > 0
- Test MSE ≥ Irreducible Error



Bias-Variance Trade off is required