PES UNIVERSITY ONLINE

DATA ANALYTICS

Unit 2: Regression – Multiple and Multivaraiate

Mamatha.H.R

Department of Computer Science and Engineering

Transformations

Transformation is a process of deriving new dependent and/or independent variables to identify the correct functional form of the regression model

Transformation in MLR is used to address the following issues:

- Poor fit (low R^2 value).
- Patten in residual analysis indicating potential non-linear relationship between the dependent and independent variable

  For example, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ is used for developing the model

  instead of $\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, resulting in a clear pattern in residual plot
- Residuals do not follow a normal distribution
- Residuals are not homoscedastic

Example

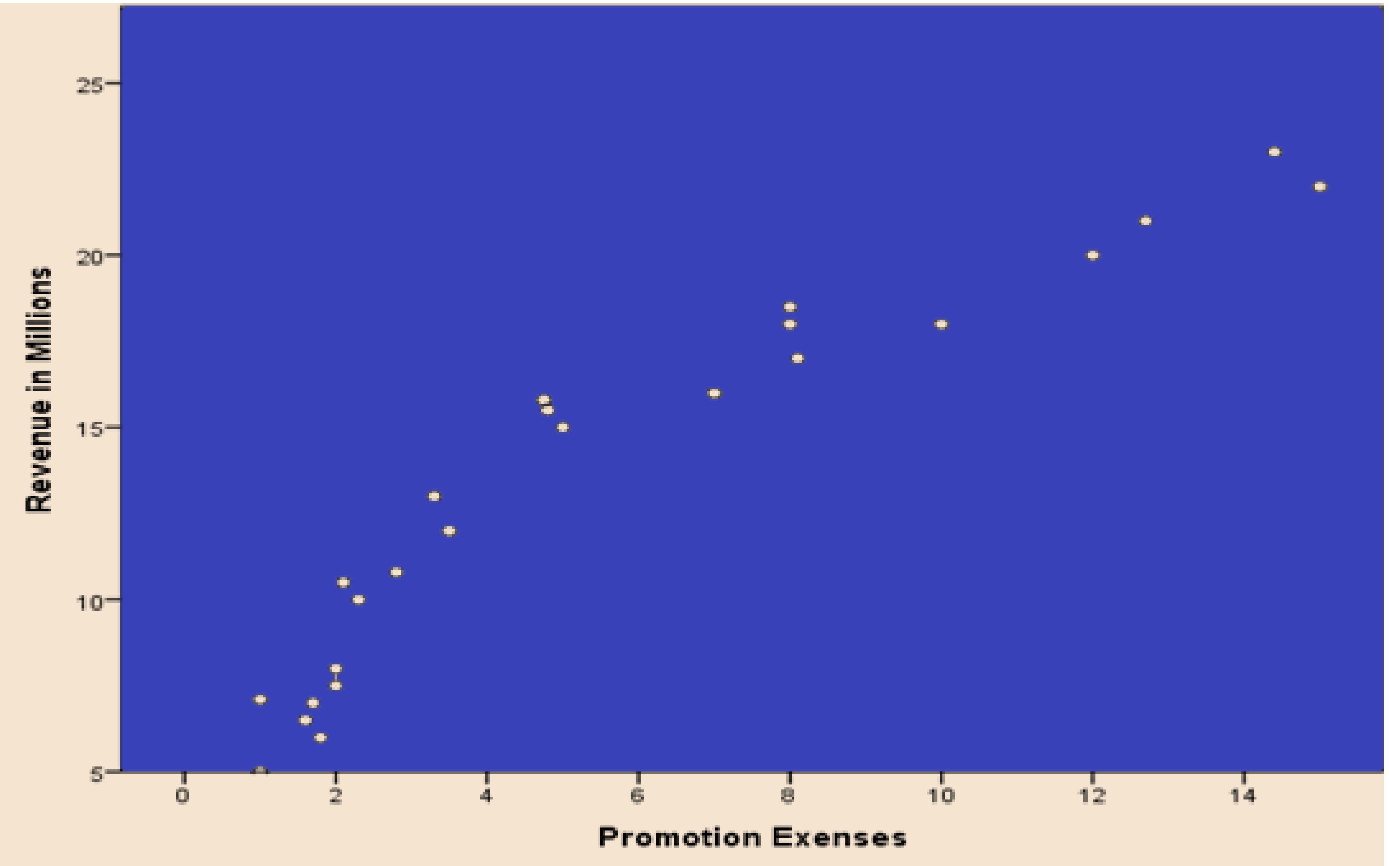
Table shows the data on revenue generated (in million of rupees) from a product and the promotion expenses (in million of rupees). Develop a regression model

S. No.	Revenue in Millions	Promotion Expenses	S. No.	Revenue in Millions	Promotion Expenses
1	5	1	13	16	7
2	6	1.8	14	17	8.1
3	6.5	1.6	15	18	8
4	7	1.7	16	18	10
5	7.5	2	17	18.5	8
6	8	2	18	21	12.7
7	10	2.3	19	20	12
8	10.8	2.8	20	22	15
9	12	3.5	21	23	14.4
10	13	3.3	22	7.1	1
11	15.5	4.8	23	10.5	2.1
12	15	5	24	15.8	4.75

Motivating Transformations

Let Y = Revenue Generated and X = Promotion Expenses

The scatter plot between Y and X for the data in Table is shown in Figure. It is clear from the scatter plot that the relationship between X and Y is not linear; it looks more like a logarithmic function.

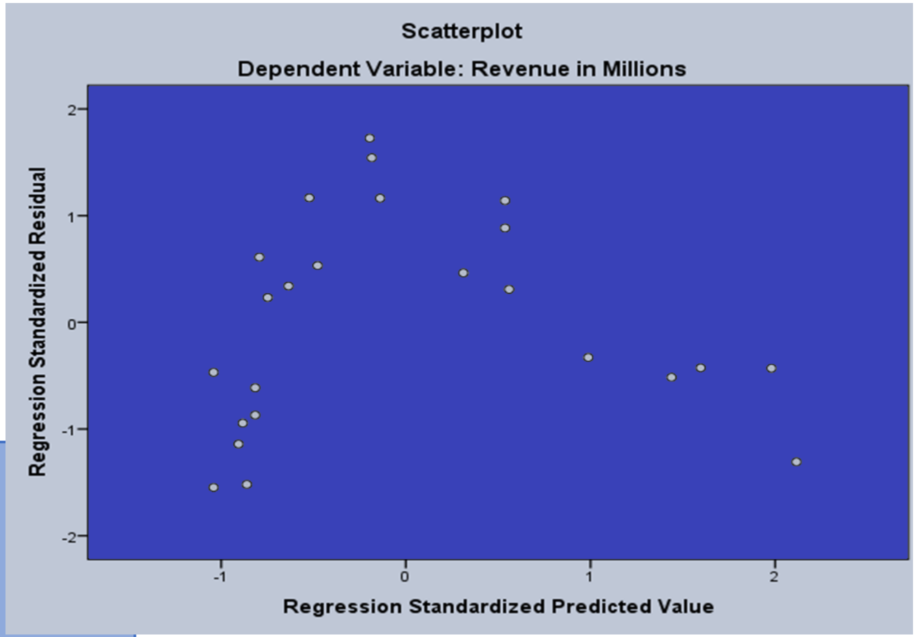


Pre-Transformation: Pattern in Residuals

Consider the function $Y = \beta_0 + \beta_1 X$. The output for this regression is shown below. There is a clear increasing and decreasing pattern in Figure indicating non-linear relationship between X and Y.

Model Summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.940	0.883	0.878	1.946



Coefficients

		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	T	Sig.
1	(Constant)	6.831	0.650		10.516	0.000
	Promotion Expenses	1.181	0.091	0.940	12.911	0.000

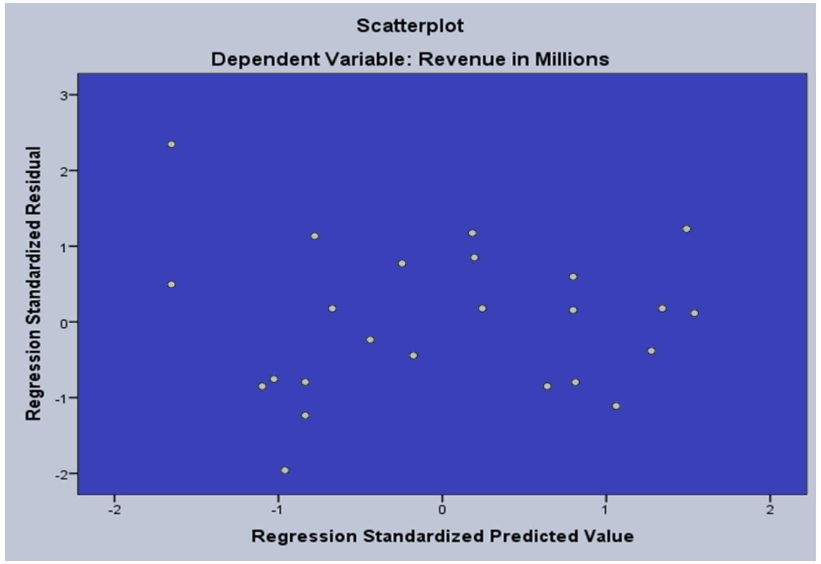
Post Transformation: No Pattern in Residuals

Since there is a pattern in the residual plot, we cannot accept the linear model ($Y = \beta_0 + \beta_1 X$).

Next we try the model $Y = \beta_0 + \beta_1 \ln(X)$. The SPSS output for $Y = \beta_0 + \beta_1\ln(X)$ and the residual plot are shown.

Model Summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.980	0.960	0.959	1.134

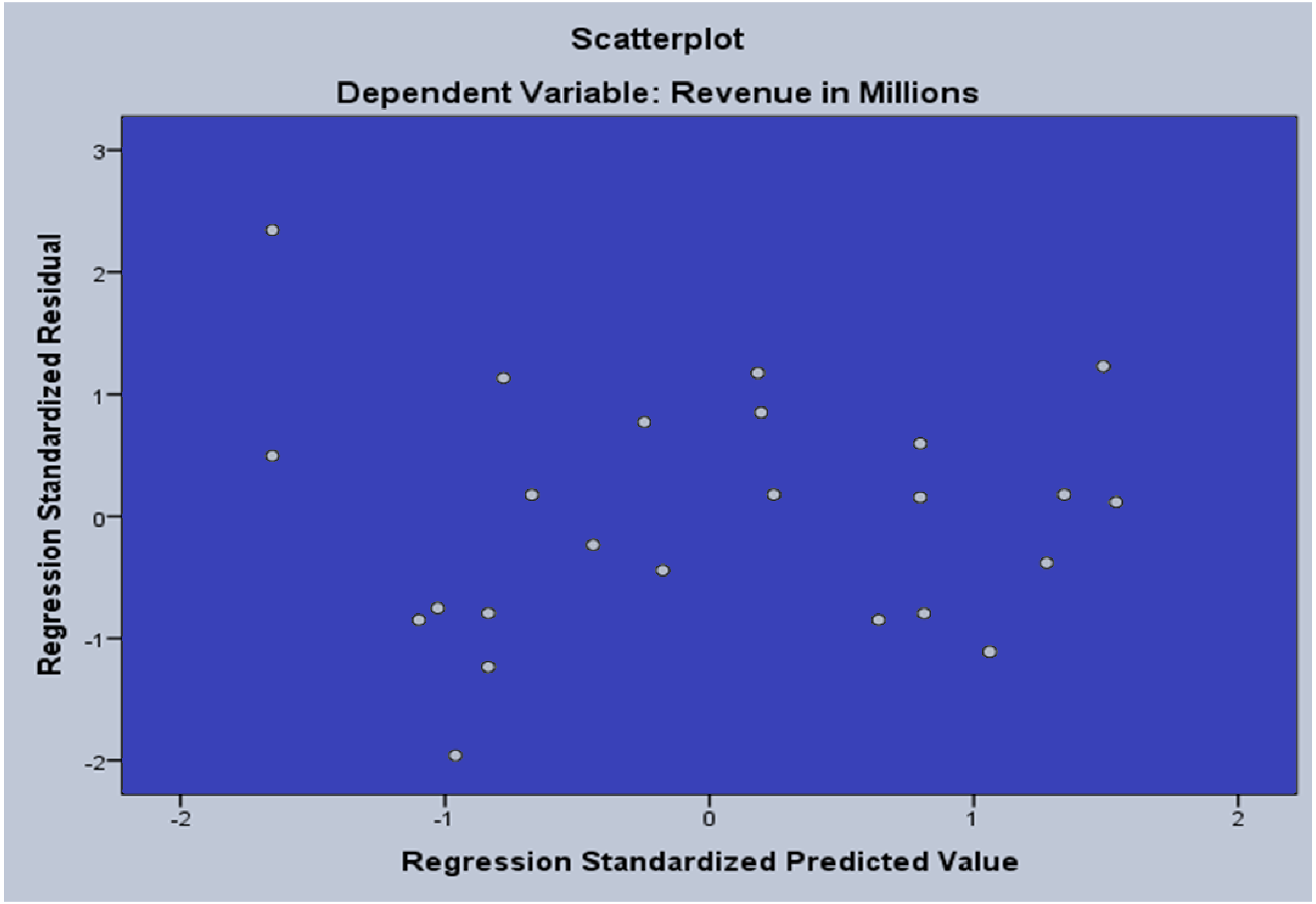


Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.439	0.454		9.771	0.000
	ln (X)	6.436	0.279	0.980	23.095	0.000

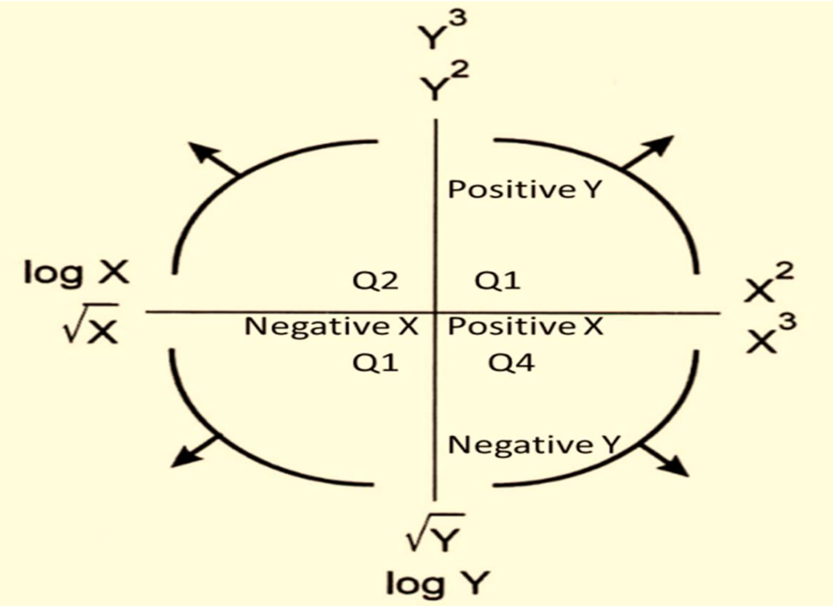
Note that for the model $Y = \beta_0 + \beta_1\ln(X)$, the R^2-value is 0.96 whereas the R^2-value for the model $Y = \beta_0 + \beta_1 X$ is 0.883. Most important, there is no obvious pattern in the residual plot of the model $Y = \beta_0 + \beta_1\ln(X)$. The model $Y = \beta_0 + \beta_1\ln(X)$ is preferred over the model $Y = \beta_0 + \beta_1 X$.

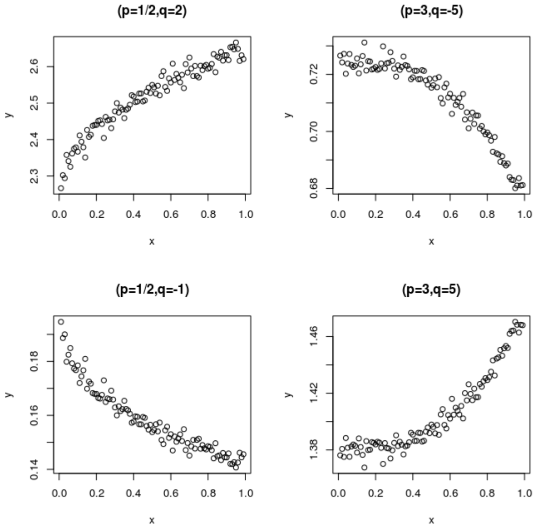
Residual plot for the model Y = $\beta_0 + \beta_1 \ln(X)$.



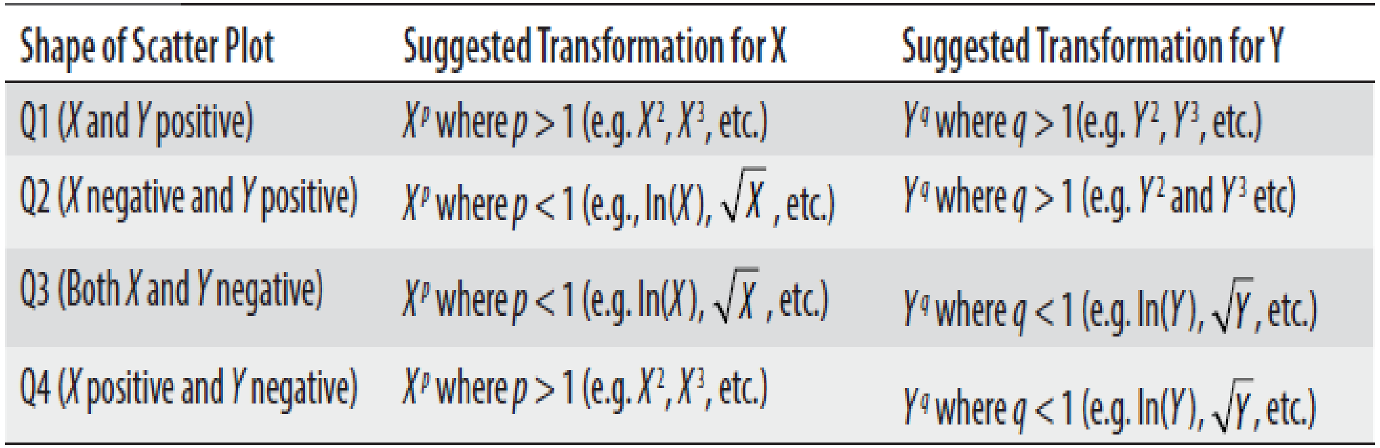
Tukey and Mosteller’s Bulging Rule for Transformation

- An easier way of identifying an appropriate transformation was provided by Mosteller and Tukey (1977), popularly known as Tukey’s Bulging Rule.
- To apply Tukey’s Bulging Rule we need to look at the pattern in the scatter plot between the dependent and independent variable.





https://freakonometrics.hypotheses.org/14967



Shape of Scatter Plot	Suggested Transformation for X	Suggested Transformation for Y
Q1 (*X* and *Y* positive)	X^p where $p > 1$ (e.g. X^2, X^3, etc.)	Y^q where $q > 1$(e.g. Y^2, Y^3, etc.)
Q2 (*X* negative and *Y* positive)	X^p where $p < 1$ (e.g., ln(*X*), $\sqrt{X}$, etc.)	Y^q where $q > 1$ (e.g. Y^2 and Y^3 etc)
Q3 (Both *X* and *Y* negative)	X^p where $p < 1$ (e.g. ln(*X*), $\sqrt{X}$, etc.)	Y^q where $q < 1$ (e.g. ln(*Y*), $\sqrt{Y}$, etc.)
Q4 (*X* positive and *Y* negative)	X^p where $p > 1$ (e.g. X^2, X^3, etc.)	Y^q where $q < 1$ (e.g. ln(*Y*), $\sqrt{Y}$, etc.)

MvLR Model: Scalar Form

Multivariate Regression (MvLR): Predict multiple dependent variables using multiple independent variables

The multivariate (multiple) linear regression model has the form

$$y_{ik} = b_{0k} + \sum_{j=1}^{p} b_{jk} x_{ij} + e_{ik}$$

for $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$ where

- $y_{ik} \in \mathbb{R}$ is the k-th real-valued response for the i-th observation
- $b_{0k} \in \mathbb{R}$ is the regression intercept for k-th response
- $b_{jk} \in \mathbb{R}$ is the j-th predictor's regression slope for k-th response
- $x_{ij} \in \mathbb{R}$ is the j-th predictor for the i-th observation
- $(e_{i1}, \ldots, e_{im}) \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}_m, \boldsymbol{\Sigma})$ is a multivariate Gaussian error vector

MvLR Model: Assumptions

The fundamental assumptions of the MLR model are:

1. Relationship between X_j and Y_k is linear (given other predictors)
2. x_{ij} and y_{ik} are observed random variables (known constants)
3. $(e_{i1}, \ldots, e_{im}) \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}_m, \boldsymbol{\Sigma})$ is an unobserved random vector
4. $\mathbf{b}_k = (b_{0k}, b_{1k}, \ldots, b_{pk})'$ for $k \in \{1, \ldots, m\}$ are unknown constants
5. $(y_{ik} | x_{i1}, \ldots, x_{ip}) \sim \mathrm{N}(b_{0k} + \sum_{j=1}^{p} b_{jk} x_{ij}, \sigma_{kk})$ for each $k \in \{1, \ldots, m\}$
   note: homogeneity of variance for each response

Note: b_{jk} is expected increase in Y_k for 1-unit increase in X_j with all other predictor variables held constant

MLR Model: Matrix Form

The multivariate multiple linear regression model has the form

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

where

- $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_m] \in \mathbb{R}^{n \times m}$ is the $n \times m$ response matrix
  - $\mathbf{y}_k = (y_{1k}, \dots, y_{nk})' \in \mathbb{R}^n$ is k-th response vector $(n \times 1)$
- $\mathbf{X} = [\mathbf{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times (p+1)}$ is the $n \times (p+1)$ design matrix
  - $\mathbf{1}_n$ is an $n \times 1$ vector of ones
  - $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})' \in \mathbb{R}^n$ is j-th predictor vector $(n \times 1)$
- $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_m] \in \mathbb{R}^{(p+1) \times m}$ is $(p+1) \times m$ matrix of coefficients
  - $\mathbf{b}_k = (b_{0k}, b_{1k}, \dots, b_{pk})' \in \mathbb{R}^{p+1}$ is k-th coefficient vector $(p + 1 \times 1)$
- $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_m] \in \mathbb{R}^{n \times m}$ is the $n \times m$ error matrix
  - $\mathbf{e}_k = (e_{1k}, \dots, e_{nk})' \in \mathbb{R}^n$ is k-th error vector $(n \times 1)$

MLR Model: Matrix Form

Matrix form writes MLR model for all *nm* points simultaneously

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

$$\begin{pmatrix} y_{11} & \cdots & y_{1m} \\ y_{21} & \cdots & y_{2m} \\ y_{31} & \cdots & y_{3m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} b_{01} & \cdots & b_{0m} \\ b_{11} & \cdots & b_{1m} \\ b_{21} & \cdots & b_{2m} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pm} \end{pmatrix} + \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ e_{21} & \cdots & e_{2m} \\ e_{31} & \cdots & e_{3m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix}$$

Fitted Values and Residuals

SCALAR FORM:

Fitted values are given by

$$\hat{y}_{ik} = \hat{b}_{0k} + \sum_{j=1}^{p} \hat{b}_{jk} x_{ij}$$

and residuals are given by

$$\hat{e}_{ik} = y_{ik} - \hat{y}_{ik}$$

MATRIX FORM:

Fitted values are given by

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$$

and residuals are given by

$$\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Hat Matrix

Note that we can write the fitted values as

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\mathbf{B}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{H}\mathbf{Y}\end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix.

$\mathbf{H}$ is a symmetric and idempotent matrix: $\mathbf{H}\mathbf{H} = \mathbf{H}$

$\mathbf{H}$ projects $\mathbf{y}_k$ onto the column space of $\mathbf{X}$ for $k \in \{1, \ldots, m\}$.

DATA ANALYTICS

Unit 2: Other forms of regression
Ridge, lasso and polynomial
Nonlinear regression

Mamatha H R, Gowri Srinivasa

Department of Computer Science and Engineering

Bias-Variance Trade-Off in Multiple Regression

The simple linear regression model, in which you aim at predicting n observations of the response variable, Y, with a linear combination of m predictor variables, X, and a normally distributed error term with variance σ^2:

$$Y = X\beta + \varepsilon,$$

$$\varepsilon \sim N(0, \sigma^2).$$

The true parameters, β, are not known we have to estimate them from the sample. In the Ordinary Least Squares (OLS) approach, we estimate them as β^ in such a way, that the sum of squares of residuals is as small as possible.

Bias-Variance Trade-Off in Multiple Regression

In other words, we minimize the following loss function:

$$L_{OLS}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 = ||y - X\hat{\beta}||^2$$

In order to obtain the infamous OLS parameter estimates,

$$\hat{\beta}_{OLS} = (X'X)^{-1}(X'Y).$$

Bias-Variance Trade-Off in Multiple Regression

In statistics, there are two critical characteristics of estimators to be considered: the bias and the variance.

The bias is the difference between the true population parameter and the expected estimator: $Bias(\hat{\beta}_{OLS}) = E(\hat{\beta}_{OLS}) - \beta.$

It measures the accuracy of the estimates. Variance, on the other hand, measures the spread, or uncertainty, in these estimates. It is given by

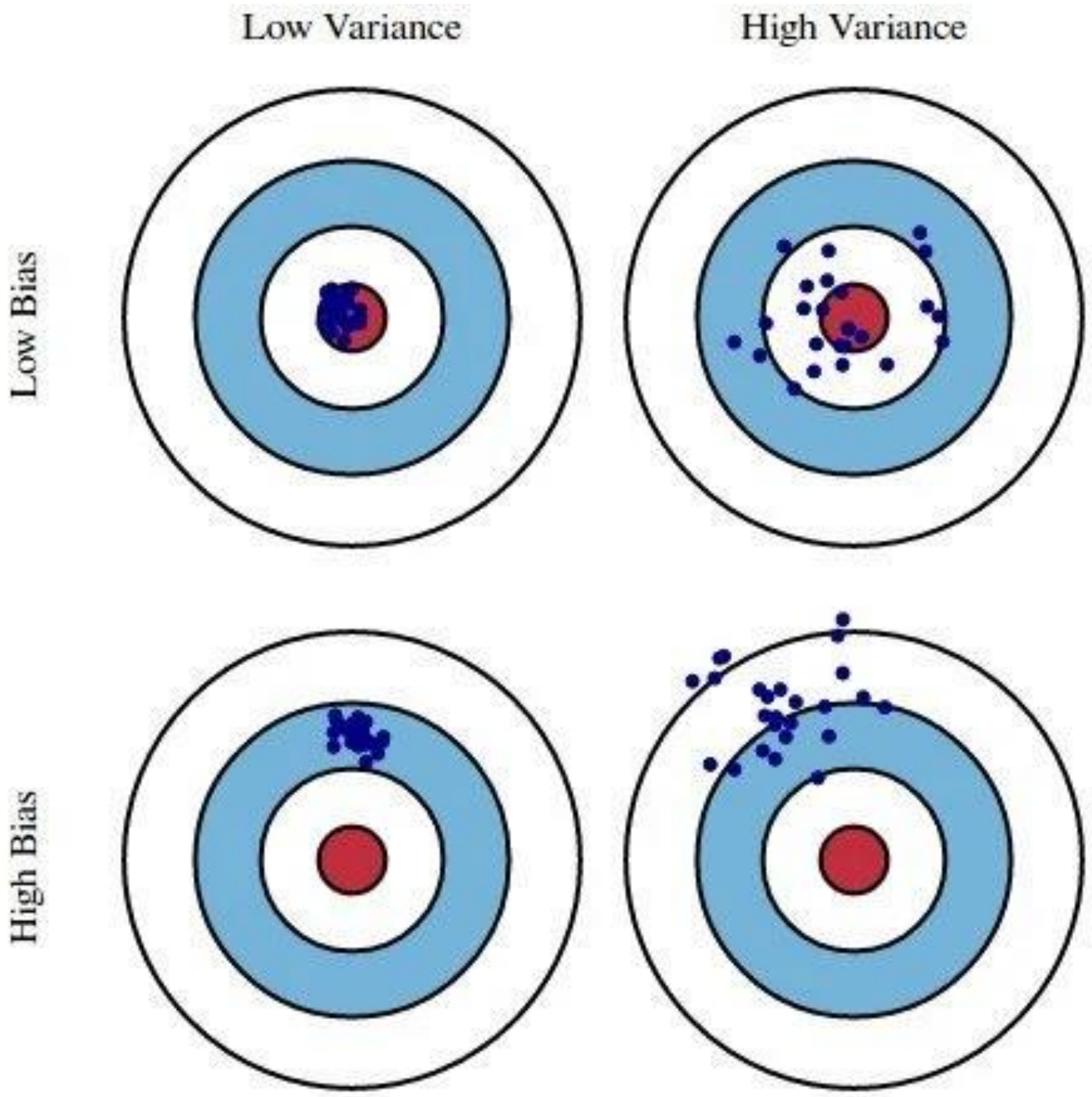
$$Var(\hat{\beta}_{OLS}) = \sigma^2 (X'X)^{-1},$$

where the unknown error variance σ^2 can be estimated from the residuals as

$$\hat{\sigma}^2 = \frac{e'e}{n-m},$$

$$e = y - X\hat{\beta}.$$

Illustration of Bias and Variance



Imagine the bull's-eye is the true population parameter that we are estimating, β, and the shots at it are the values of our estimates resulting from four different estimators - low bias and variance, high bias and variance, and the combinations.

Bias-Variance Trade-Off in Multiple Regression

Both the bias and the variance are desired to be low, as large values result in poor predictions from the model.

- In fact, the model's error can be decomposed into three parts:
  - error resulting from a large variance,
  - error resulting from significant bias,
  - and the remainder - the unexplainable part.

$$E(e) = (E(X\hat{\beta}) - X\beta)^2 + E(X\hat{\beta} - E(X\hat{\beta}))^2 + \sigma^2 =$$
$$Bias^2 + Variance + \sigma^2$$

Bias-Variance Trade-Off in Multiple Regression

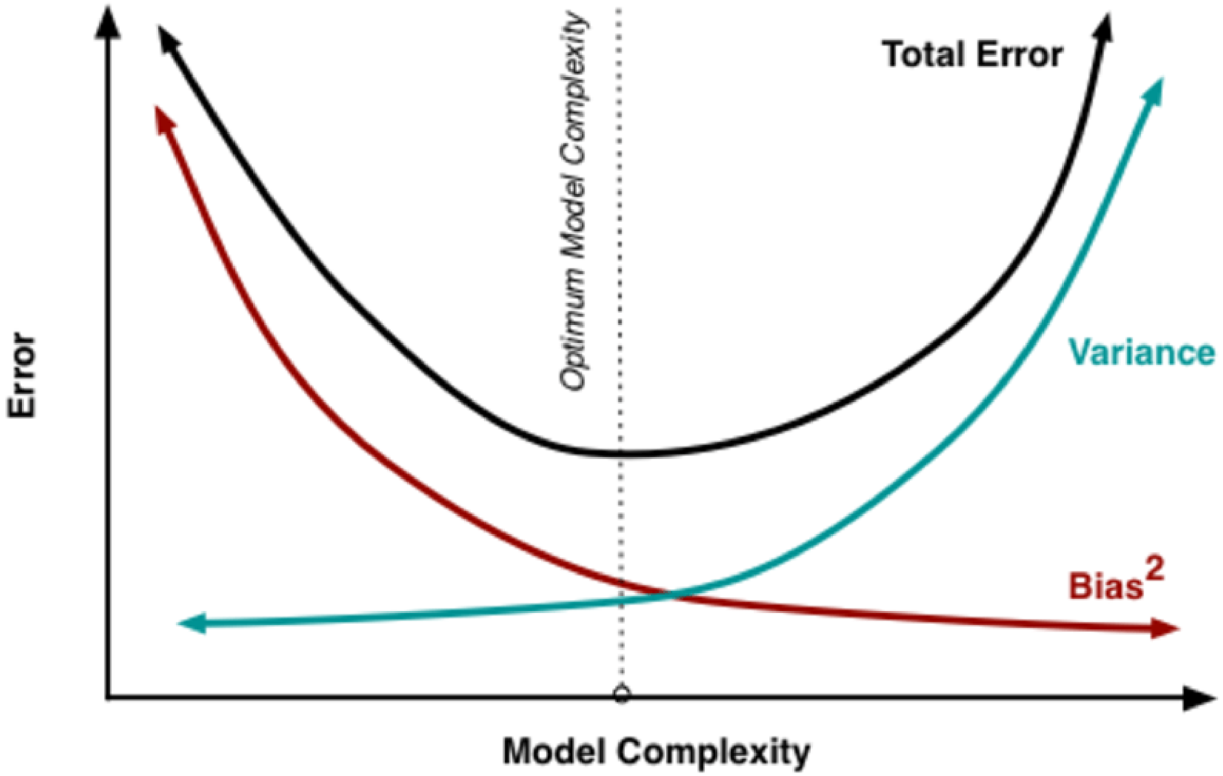
The OLS estimator has the desired property of being unbiased. However, it can have a huge variance.

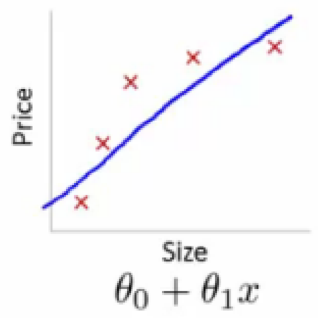
Specifically, this happens when:

- The predictor variables are highly correlated with each other
- There are many predictors.
- If m approaches n, the variance approaches infinity $\hat{\sigma}^2 = \frac{e'e}{n-m},$

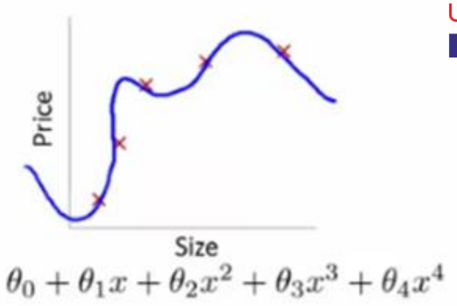
- The general solution to this is:
  reduce variance at the cost of introducing some bias
- This approach is called regularization and is almost always beneficial for the predictive performance of the model

Bias-Variance Trade-Off in Multiple Regression

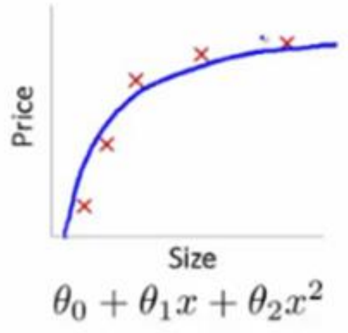




High bias (underfit)



High variance (overfit)



“Just right”

https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/

What do Lasso and Ridge Regression do?

To decrease the model complexity, that is the number of predictors.

We could use the forward or backward selection for this, but that way we would not be able to tell anything about the removed variables' effect on the response.

Removing predictors from the model can be seen as settings their coefficients to zero (Lasso).

Instead of forcing them to be exactly zero, let's penalize them if they are too far from zero, thus enforcing them to be small in a continuous way (Ridge).

This way, we decrease model complexity while keeping all variables in the model.

Lasso Regression

- LASSO: Least absolute shrikange and selection
- Assumptions same as linear regression, normality not assumed
- Uses L_1 norm or the 'absolute value' of coefficients scaled by shrinkage
- λ is a tunable parameter
- Lasso tends to zero out smaller (unimportant) coefficients (and helps with feature selection)

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

Limitations:

• In small-n-large-p dataset the LASSO selects at most n variables before it saturates.

• If there are grouped variables (highly correlated between each other) LASSO tends to select one variable from each group ignoring the others

Ridge Regression

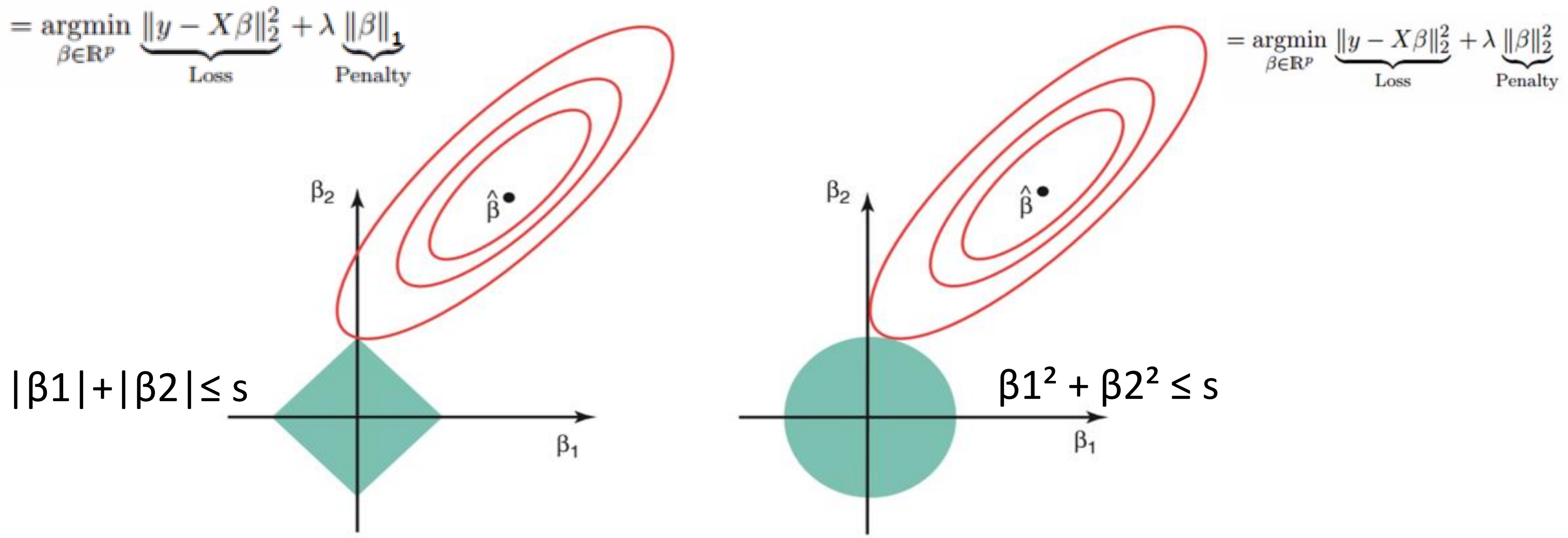
- Assumptions same as linear regression, normality not assumed
- A shrinkage term is added to the objective (SSE) function
- λ is a tunable parameter; penalizes flexibility of the model
- We shrink the estimated association of each variable
- $\lambda=0$ has no effect and as $\lambda \to \infty$ and ridge regression coefficient estimates approach 0

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

- Coefficients produced by OLS are scale invariant but that is not the case with Ridge Regression, so we must remember to scale the input

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

Lasso Vs Ridge



$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

$|\beta 1|+|\beta 2| \leq s$

$\beta 1^2 + \beta 2^2 \leq s$

- Correlated variables have similar weights using Ridge and whereas one is high and the other(s) nearly zero with Lasso
- Interpretability: Lasso zeros out unimportant coefficients and hence performs feature selection whereas ridge gives a small weight but includes them all
- Both achieve reduction in variance without increase in bias

https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a

RIDGE REGRESSION

- Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated).
- In this phenomenon, one predicted value in multiple regression models is linearly predicted with others to attain a certain level of accuracy.
- The concept multicollinearity occurs when there are high correlations between more than two predicted variables.
- In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value.
- By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Choice of Regularization Parameter

Question : how much bias are we willing to accept in order to decrease the variance? Or: what is the optimal value for λ?

- Choose λ such that some information criterion, e.g., AIC or BIC, is the smallest. approach emphasizes the model's fit to the data
- This approach boils down to estimating the model with many different values for λ and choosing the one that minimizes the Akaike or Bayesian Information Criterion:

$$AIC_{ridge} = nlog(e'e) + 2df_{ridge},$$
$$BIC_{ridge} = nlog(e'e) + 2df_{ridge}\,log(n),$$

where df_{ridge} is the number of degrees of freedom

- A more machine learning-like approach is to perform cross-validation and select the value of λ that minimizes the cross-validated sum of squared residuals.

ElasticNet Regression: Combining L1 and L2 norms

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

- Inherits Ridge’s stability under rotation
- It encourages group effect in case of highly correlated variables
- There are no limitations on the number of selected variables
- It can suffer with double shrinkage

https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

References

https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net

https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-

https://www.statisticshowto.com/lasso-regression

https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf

Polynomial Function: Definition

Reminder: a polynomial function has the form

$$\begin{aligned} f(x) &= a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n \\ &= \sum_{j=0}^{n} a_j x^j \end{aligned}$$

where $a_j \in \mathbb{R}$ are the coefficients and x is the indeterminate (variable).

Note: x^j is the j-th order polynomial term

- x is first order term, x^2 is second order term, etc.
- The degree of a polynomial is the highest order term

Model Form (scalar)

The polynomial regression model has the form

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_i^j + e_i$$

for $i \in \{1, \ldots, n\}$ where

- $y_i \in \mathbb{R}$ is the real-valued response for the i-th observation
- $b_0 \in \mathbb{R}$ is the regression intercept
- $b_j \in \mathbb{R}$ is the regression slope for the j-th degree polynomial
- $x_i \in \mathbb{R}$ is the predictor for the i-th observation
- $e_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is a Gaussian error term

Model Form (matrix)

The polynomial regression model has the form

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ 1 & x_3 & x_3^2 & \cdots & x_3^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{pmatrix}$$

Note that this is still a linear model, even though we have polynomial terms in the design matrix.

PR Model Assumptions (scalar)

The fundamental assumptions of the PR model are:

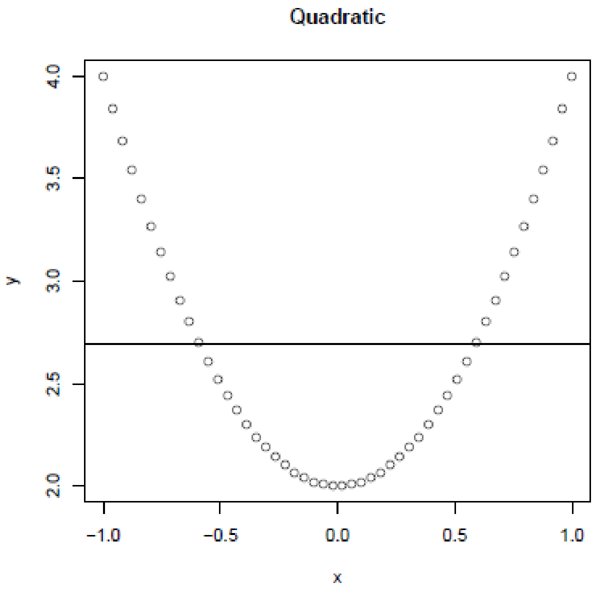
1. Relationship between X and Y is polynomial
2. x_i and y_i are observed random variables (known constants)
3. $e_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is an unobserved random variable
4. $b_0, b_1, \ldots, b_p$ are unknown constants
5. $(y_i | x_i) \overset{\text{ind}}{\sim} \mathrm{N}(b_0 + \sum_{j=1}^{p} b_j x_i^j, \sigma^2)$
   note: homogeneity of variance

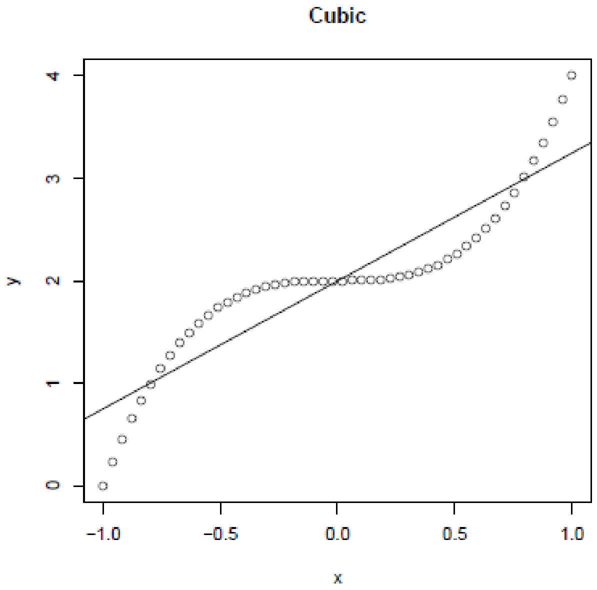
Note: focus is estimation of the polynomial curve.

Polynomial Function: Simple Regression

```
> x=seq(-1,1,length=50)
> y=2+2*(x^2)
> plot(x,y,main="Quadratic")
> qmod=lm(y~x)
> abline(qmod)
```

```
> x=seq(-1,1,length=50)
> y=2+2*(x^3)
> plot(x,y,main="Cubic")
> cmod=lm(y~x)
> abline(cmod)
```





Polynomial Regression: Properties

Some important properties of the PR model include:

1. Need $n > p$ to fit the polynomial regression model
2. Setting $p = 1$ produces simple linear regression
3. Setting $p = 2$ is quadratic polynomial regression
4. Setting $p = 3$ is cubic polynomial regression
5. Rarely set $p > 3$; use cubic spline instead

Multicollinearity: Problem

Note that x_i, x_i^2, x_i^3, etc. can be highly correlated with one another, which introduces multicollinearity problem.

```
> set.seed(123)
> x = runif(100)*2
> X = cbind(x, xsq=x^2, xcu=x^3)
> cor(X)
            x       xsq       xcu
x   1.0000000 0.9703084 0.9210726
xsq 0.9703084 1.0000000 0.9866033
xcu 0.9210726 0.9866033 1.0000000
```

Multicollinearity: Partial Solution

You could mean-center the x_i terms to reduce multicollinearity.

```
> set.seed(123)
> x = runif(100)*2
> x = x - mean(x)
> X = cbind(x, xsq=x^2, xcu=x^3)
> cor(X)
             x        xsq        xcu
x   1.00000000 0.03854803 0.91479660
xsq 0.03854803 1.00000000 0.04400704
xcu 0.91479660 0.04400704 1.00000000
```

But this doesn’t fully solve our problem. . .

Orthogonal Polynomials: Definition

To deal with multicollinearity, define the set of variables

$$
\begin{aligned}
z_0 &= a_0 \\
z_1 &= a_1 + b_1 x \\
z_2 &= a_2 + b_2 x + c_2 x^2 \\
z_3 &= a_3 + b_3 x + c_3 x^2 + d_3 x^3
\end{aligned}
$$

where the coefficients are chosen so that $z_j' z_k = 0$ for all $j \neq k$.

The transformed z_j variables are called orthogonal polynomials.

How to detect non linearity ?

1. Theory – in many sciences we have theories about nonlinear relations within some phenomenon.

2. Scatterplot – when looking at the plot you can see that data points are not linear or even not nearly linear.

3. Seasonality in data – i.e. in agriculture, building industry we often have seasonality within the data.

4. Estimated model does not fit the data well or does not fit it at all; the estimated β's are not significant – this might suggest nonlinearity.

5. Can often do incremental F tests

NON LINEAR REGRESSION

Central idea of non-linear regression: same as linear regression, just with non-linear features

Two ways to construct non-linear features

1. Explicitly (construct actual feature vector)
2. Implicitly (using kernels)

$$\text{E.g. } \phi(x_i) = \begin{bmatrix} x_i^2 \\ x_i \\ 1 \end{bmatrix}$$

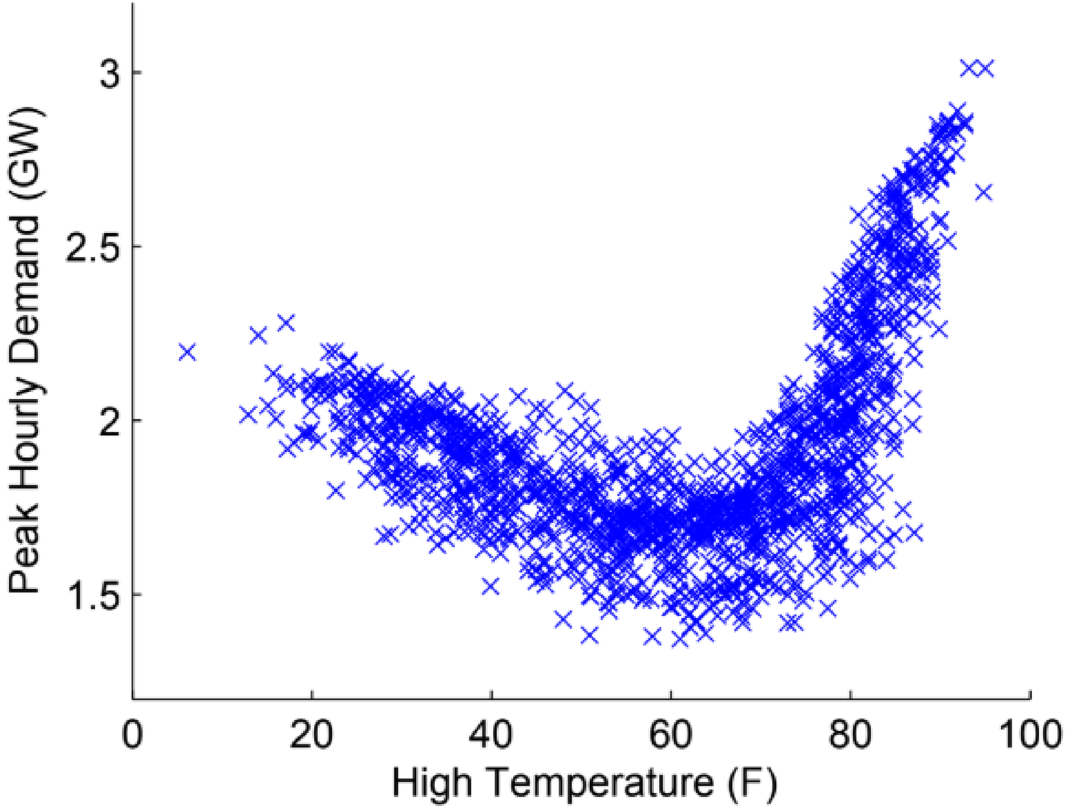
NON LINEAR REGRESSION

Some popular nonlinear regression models:

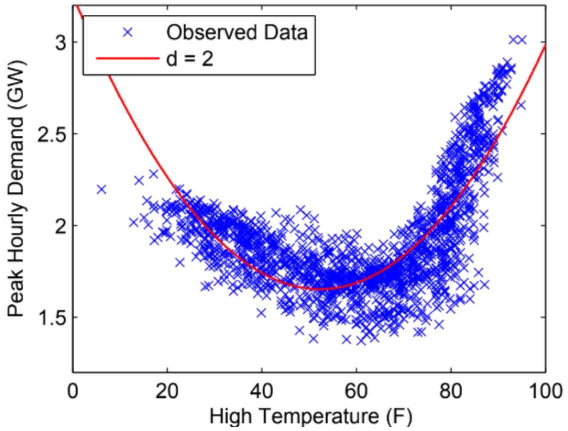
1. Exponential model: $(y = ae^{bx})$
2. Power model: $(y = ax^{b})$
3. Saturation growth model: $\left(y = \dfrac{ax}{b+x}\right)$

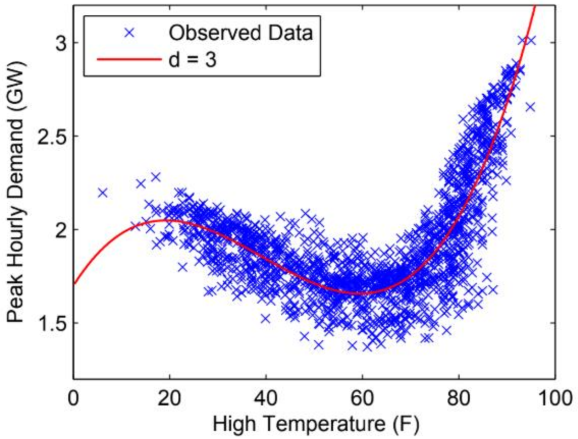
NON LINEAR REGRESSION

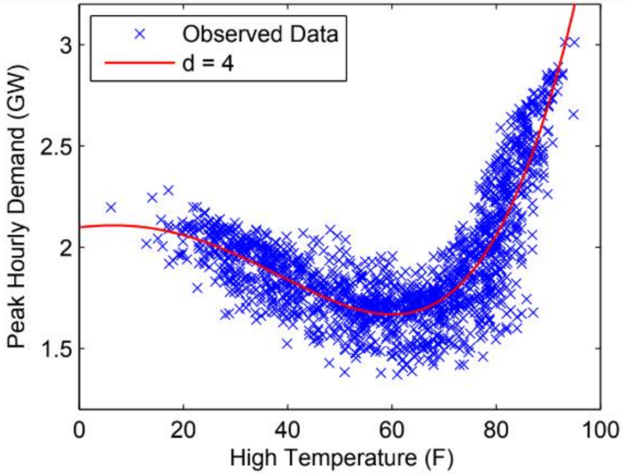
Non-linear regression



NON LINEAR REGRESSION







Degree 4 polynomial

Why use non-linear regression?

Transformation is necessary to obtain variance homogeneity, but transformation destroys linearity.

Linearity does not fit, and the transformation seems to destroy other parts of the model assumptions, e.g. the assumption of variance homogeneity.

Theoretical knowledge indicates that the proper relation is intrinsically non-linear.

Interest is in functions of the parameters that do not enter linearly in the model

Some questions asked after class

- **Is there proof we are reducing total error if we add bias?**

- While there is no 'proof' other than our empirical understanding of the total error (which is convex and) comprises the bias, variance and 'unexplained' residuals (or 'noise)

  We can rephrase this question as: is there anyway to detect high bias or high variance? The answer is 'yes' and there are ways of remedying these problems as well:

- **Detection of high variance (Regime #1)**

  1.Training error is much lower than test error

  2.Training error is lower than some desired tolerance 't'

  3.Test error is above 't'

  Remedies

  a. Add more training data

  b. Reduce model complexity -- complex models are prone to high variance

  c. Bagging (will be taught in the course on Machine Intelligence)

- **Detection of high bias in the model (Regime #2)**

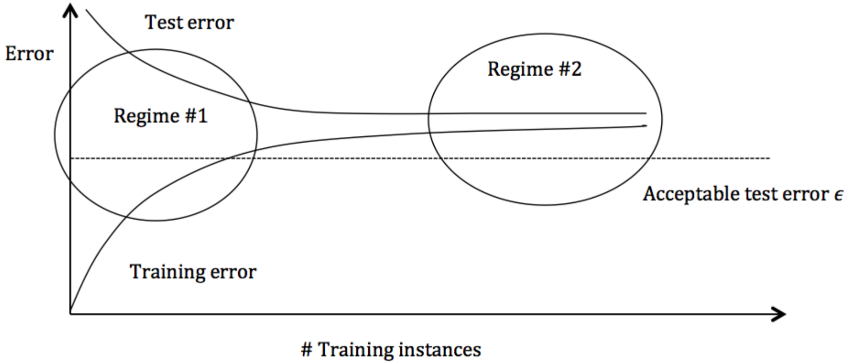
  1. Training error is higher than a desired tolerance 't'

  Remedies

  a. Use more complex model (e.g. kernelize, use non-linear models)

  b. Add features

  c. Boosting (will be covered later in the course)



https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html

- **How do we know the error function is indeed convex?**

  The error function (SSE) is a quadratic function in one variable and hence, convex

  https://towardsdatascience.com/understanding-convexity-why-gradient-descent-works-for-linear-regression-aaf763308708

  https://www.princeton.edu/~aaa/Public/Teaching/ORF523/S16/ORF523_S16_Lec7_gh.pdf

Some questions asked after class

- **Can we minimize both bias and variance using Ridge/ Lasso**

- Please note: Both bias and variance cannot be arbitrarily minimized
- We are *adding* bias to minimize the overall error (i.e., reduce the high variance)

- **Can the shrinkage parameter (lambda) be positive, negative or zero?**

- In theory yes; though let us see what each option for the shrinkage value would mean:
- Zero: same as simple linear regression (bias would be low (or zero) $\Rightarrow$ variance is very high)
- Positive: we are reducing the bias; if the overall equation is to be minimized, then larger the λ, the closer β's would get to zero
- Negative: if lambda is negative, our loss can be reduced indefinitely; it means (some of the) β's$\rightarrow\infty$
- So, in practice, lambda is range limited (i.e., we insist on $\lambda>0$)

References

Additional references (for the interested student)

http://users.stat.umn.edu/~helwig/notes/polyint-Notes.pdf

http://users.stat.umn.edu/~helwig/notes/mvlr-Notes.pdf

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834