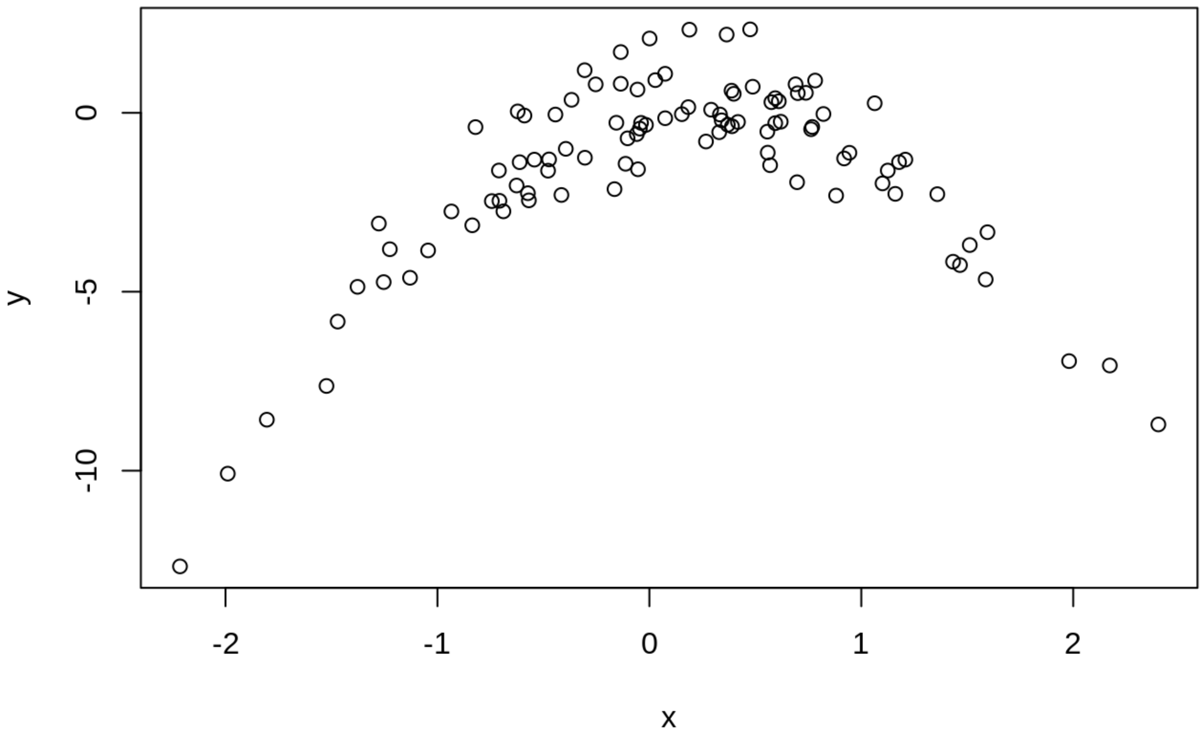
P-4-SimulatedData

Jeetendra Gan

11/5/2019

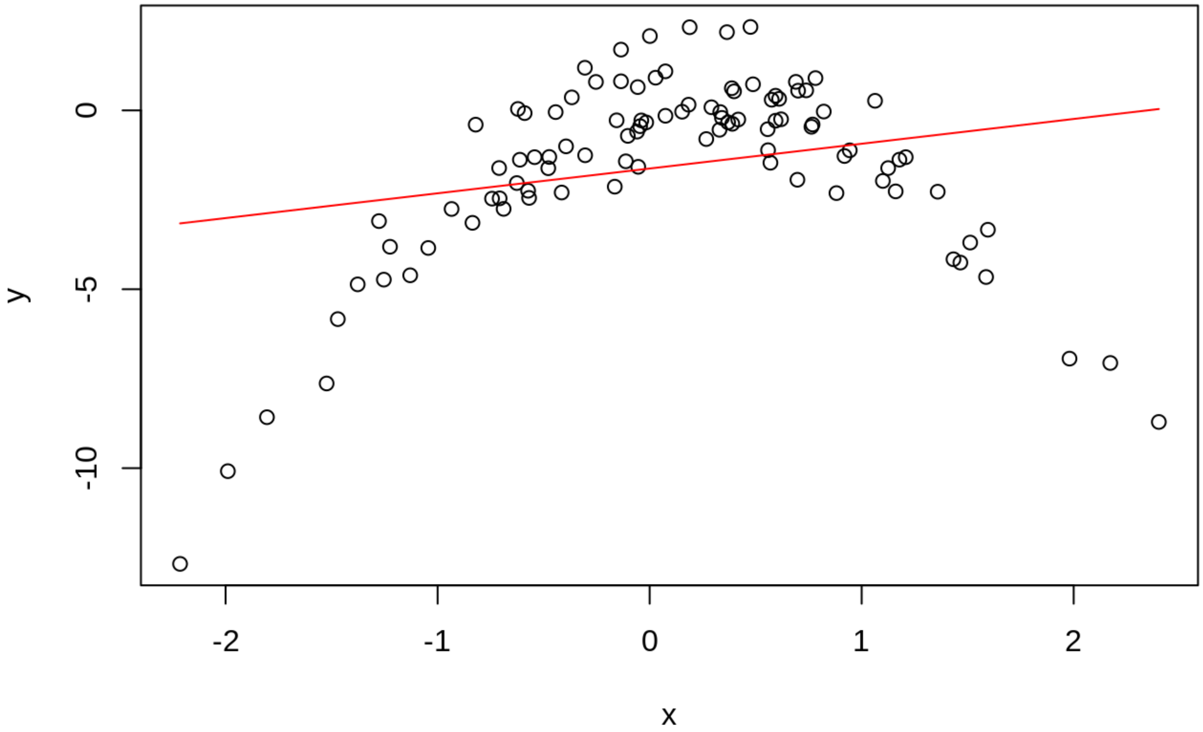
The given polynomial $y = x - 2x^2 + e$ is most equivalent to $y = b_0 + b_1x + b_2x^2$

Here is the simulated data as a graph:



LOOCV for polynomial with degree 1

Following figure shows how a linear model with degree 1 fits the data.



The coefficient estimates for the degree 1 fit are:

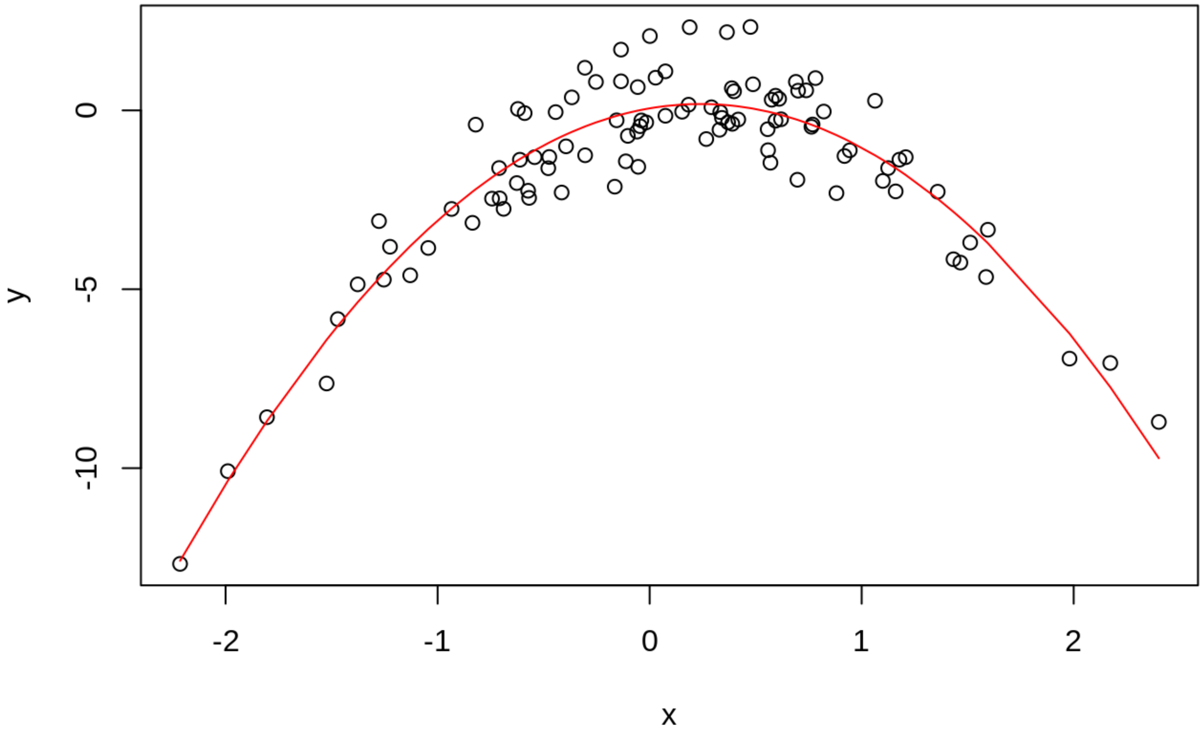
```
## (Intercept)           x
##   -1.625427    0.692497
```

The LOOCV error is as follows:

```
## [1] 7.288162
```

LOOCV for polynomial with degree 2

The fit for a polynomial with degree 2 is shown in the figure below:



The coefficient estimates are as follows:

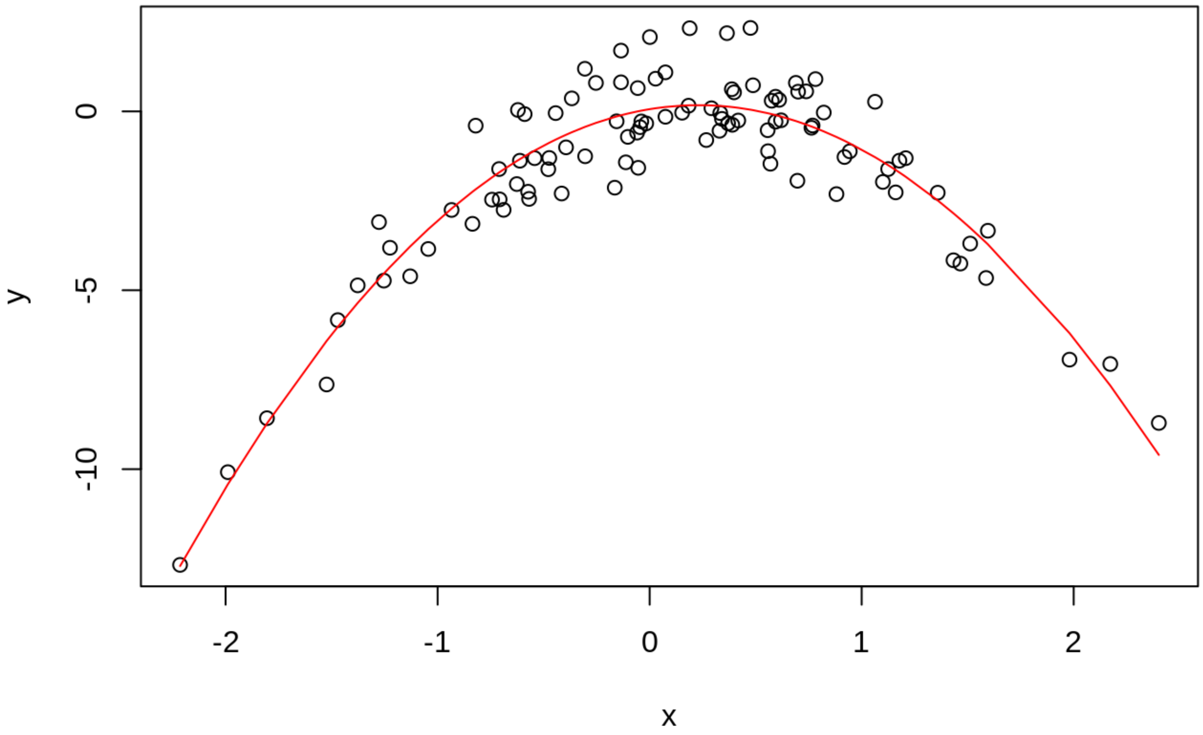
```
## (Intercept) poly(x, 2)1 poly(x, 2)2
##   -1.550023    6.188826  -23.948305
```

The LOOCV error for polynomial with degree 2 is:

```
## [1] 0.9374236
```

LOOCV for polynomial with degree 3

The fit for a polynomial with degree 3 is shown in the figure below:



Coeffienent estimates:

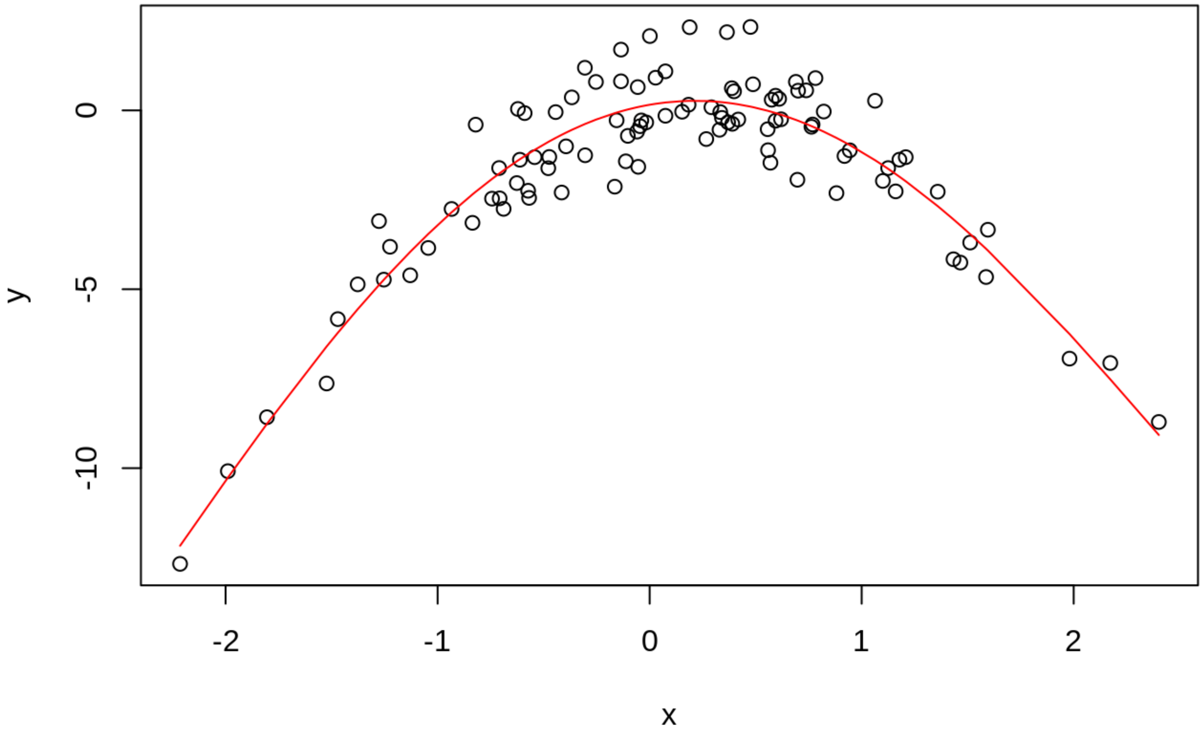
```
## (Intercept) poly(x, 3)1 poly(x, 3)2 poly(x, 3)3
##  -1.5500226   6.1888256 -23.9483049   0.2641057
```

LOOCV error:

```
## [1] 0.9566218
```

LOOCV for polynomial with degree 4

The fit for a polynomial with degree 4 is shown in the figure below:



Coeffienent estimates:

```
## (Intercept) poly(x, 4)1 poly(x, 4)2 poly(x, 4)3 poly(x, 4)4
##  -1.5500226   6.1888256 -23.9483049   0.2641057   1.2570950
```

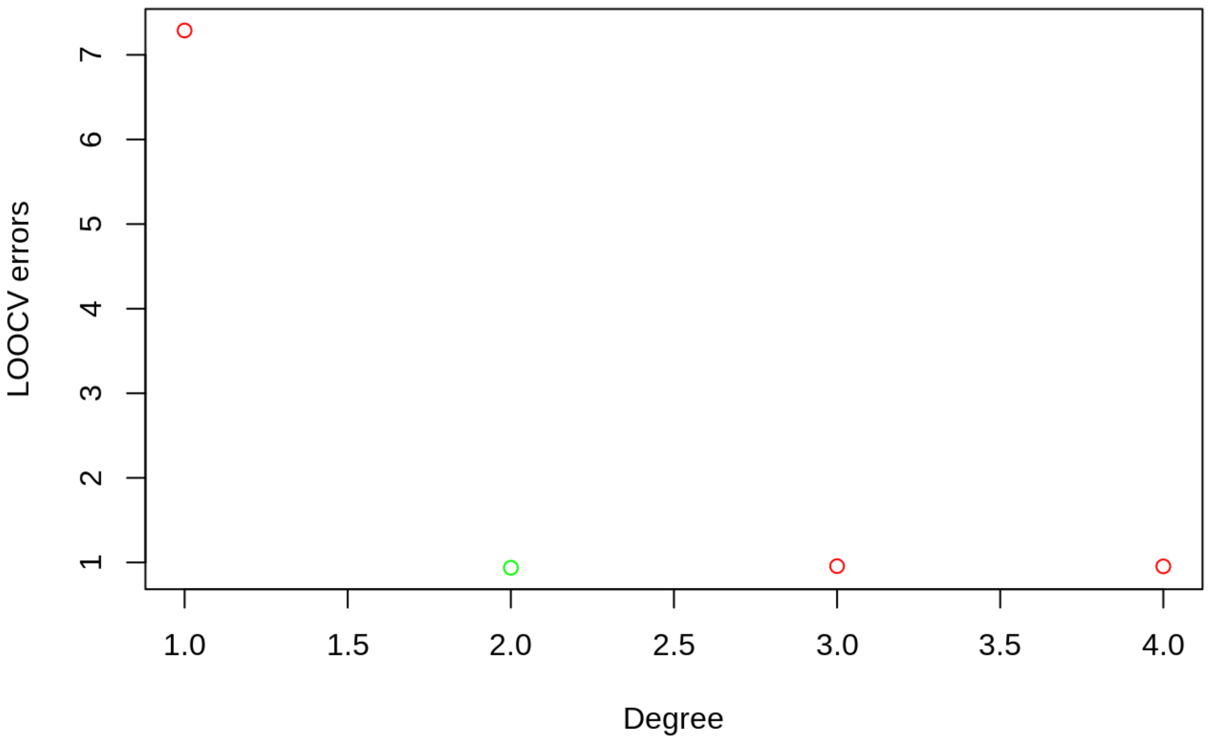
LOOCV error:

```
## [1] 0.9539049
```

LOOCV error estimates for each degree

Following are the LOOCV estimates for each degree followed by a graph marked (in green) with the least LOOCV error estimate.

```
## [1] 7.2881616 0.9374236 0.9566218 0.9539049
```



Which of the models had the smallest LOOCV error?

The model with degree 2 has the smallest LOOCV error. This was expected as the true model, $y = x - 2x^2 + e$ also has a degree 2. The polynomial with degree 1 has the heighest error which suggests that it is clearly a bad fit. The models with degree 3 and 4 will better explain the variance in the data than the one with degree 2. But, they will overfit the data and will not perform well for unseen data, hence their LOOCV error estimates are slightly greater.

Statistical significance of the coefficient estimates

Following is the summary of the fits with degree 1, 2, 3, and 4. It can be seen that the p-values for heigher order coefficients(power 3 and 4) is very less for fits with degree 3 and 4, which indicates that they are not significant. For a model with degree 1, the p-values indicate the power 1 coefficient is slightly significant. For the model with degree 2, each of the 3 coefficients are highly significant.

The results agree with the conclusion (i.e. the best model is with a degree 2) drawn from cross validation.

```
##
## Call:
## glm(formula = y ~ x, data = simData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.5161  -0.6800   0.6812   1.5491   3.8183
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.6254     0.2619  -6.205 1.31e-08 ***
## x             0.6925     0.2909   2.380   0.0192 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 6.760719)
##
##     Null deviance: 700.85  on 99  degrees of freedom
## Residual deviance: 662.55  on 98  degrees of freedom
## AIC: 478.88
##
## Number of Fisher Scoring iterations: 2
```

```
## 
## Call:
## glm(formula = y ~ poly(x, 2), data = simData)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.9650  -0.6254  -0.1288   0.5803   2.2700  
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -1.5500     0.0958  -16.18  < 2e-16 ***
## poly(x, 2)1   6.1888     0.9580    6.46 4.18e-09 ***
## poly(x, 2)2 -23.9483     0.9580  -25.00  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 0.9178258)
## 
##     Null deviance: 700.852  on 99  degrees of freedom
## Residual deviance:  89.029  on 97  degrees of freedom
## AIC: 280.17
## 
## Number of Fisher Scoring iterations: 2
```

```
## 
## Call:
## glm(formula = y ~ poly(x, 3), data = simData)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.9765  -0.6302  -0.1227   0.5545   2.2843  
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -1.55002    0.09626 -16.102  < 2e-16 ***
## poly(x, 3)1   6.18883    0.96263   6.429 4.97e-09 ***
## poly(x, 3)2 -23.94830    0.96263 -24.878  < 2e-16 ***
## poly(x, 3)3   0.26411    0.96263   0.274    0.784    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 0.9266599)
## 
##     Null deviance: 700.852  on 99  degrees of freedom
## Residual deviance:  88.959  on 96  degrees of freedom
## AIC: 282.09
## 
## Number of Fisher Scoring iterations: 2
```

```
##
## Call:
## glm(formula = y ~ poly(x, 4), data = simData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0550  -0.6212  -0.1567   0.5952   2.2267
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.55002    0.09591 -16.162  < 2e-16 ***
## poly(x, 4)1   6.18883    0.95905   6.453 4.59e-09 ***
## poly(x, 4)2 -23.94830    0.95905 -24.971  < 2e-16 ***
## poly(x, 4)3   0.26411    0.95905   0.275    0.784
## poly(x, 4)4   1.25710    0.95905   1.311    0.193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.9197797)
##
##     Null deviance: 700.852  on 99  degrees of freedom
## Residual deviance:  87.379  on 95  degrees of freedom
## AIC: 282.3
##
## Number of Fisher Scoring iterations: 2
```